-1 C 5170-1 -

Medical Care

February 1982
Vol. 20 No. 2
MDLCBD · ISSN 0025-7079

OFFICIAL JOURNAL OF THE MEDICAL CARE SECTION,
AMERICAN PUBLIC HEALTH ASSOCIATION

Physician Performance in a Prepaid Health Plan: Results of the Peer Review Program of the Health Insurance Plan of Greater New York

Episode-Based Efficiency Comparisons for Physicians and Nurse Practitioners

Measurement of Task Delegations Among Nurses by Nominal Group Process Analysis

Costs and Outcomes of Medical Intensive Care

Predicting the Outcome of Primary Care

Hospital Discharge Queues in Massachusetts

Privacy and Computerized Health Information in Community Health Centers and District Health Agencies

EPSDT Impact on Health Status: Estimates Based on Secondary Analysis of Administratively Generated Data

PUBLISHED BY
J. B. Lippincott Company · Philadelphia · Toronto

EDITORIAL BOARD

Donald C. Riedel, Ph.D.
Editor-in-Chief
University of Washington

Ian Van Duser
Editorial Assistant

S. E. BERKE, M.A.
University of Michigan

ROBERT H. BROOK, M.D., Sc.D.
University of California
Los Angeles

GORDON H. DE FRIESE, Ph.D.
University of North Carolina

EUGENE FEINGOLD, Ph.D.
University of Michigan

BRANT E. FRIES, Ph.D.
Yale University

THOMAS R. GOODKIN, R.N., P.A.
University of Oklahoma

HAROLD HUNTER, Dr. P.H.
Washington, D.C.

SHERMAN A. JAMES, Ph.D.
University of North Carolina

ANTHONY L. KOMAROFF, M.D.
Brigham and Women's
Hospital, Boston

JAMES P. LOGERFO, M.D., M.P.H.
University of Washington

NANCY MILIO, Ph.D.
University of North Carolina

DUNCAN NEUHAUSER, Ph.D.
Case Western Reserve University

JOHN E. WARE, JR., Ph.D.
The Rand Corporation

MONROE S. YERBY, M.D.
Harvard University

ALFRED YANKAUER, M.D.
University of Massachusetts
Medical School

MEDICAL CARE (ISSN 0025-7079) is published monthly by J. B. Lippincott Company, East Washington Square, Philadelphia, PA 19105. Subscription address: Journal Fulfillment Department, Lippincott/Harper, 2350 Virginia Avenue, Hagerstown, MD 21740. Subscription rates: United States and U.S. Possessions, one year, $45.00, special price U.S. students, $23.00, all other countries except Japan, one year, $55.00. Single copies $6.00. Subscription orders and correspondence concerning subscriptions in Japan should be sent to Woodbell Scope, 11-11 Shoto 2-chome, Shibuya-ku, Tokyo 150, Japan. Printed in U.S.A.
Copyright © 1982 by J. B. Lippincott Company.
Second class postage paid at Hagerstown, Maryland, and additional mailing offices.

February 1982
Volume 20, Number 2

CONTENTS

ORIGINAL ARTICLES

Physician Performance in a Prepaid Health Plan: Results of the Peer Review Program of the Health Insurance Plan of Greater New York. *Jeanne Mazzagna Deuschle, Barbara Alvarez, Donald N. Logsdon, William M. Stahl and Harry Smith, Jr.* 127

Episode-Based Efficiency Comparisons for Physicians and Nurse Practitioners. *David S. Salkever, Elizabeth A. Skinner, Donald M. Steinwachs and Harvey Katz* 143

Measurement of Task Delegations Among Nurses by Nominal Group Process Analysis. *Vandan M. Trivedi* 154

Costs and Outcomes of Medical Intensive Care. *Mark R. Chassin* 165

Predicting the Outcome of Primary Care. *Diana Dyer-Wright and Robert L. Kane* 180

Hospital Discharge Queues in Massachusetts. *Leonard W. Gruenberg and Thomas R. Willemain* 188

Privacy and Computerized Health Information in Community Health Centers and District Health Agencies. *Barry Evatt and Robb Town* 202

EPSDT Impact on Health Status: Estimates Based on Secondary Analysis of Administratively Generated Data. *Patrick H. Irwin and Rosemary Conroy-Hughes* 216

COMMUNICATION

Primary Care and Medical Education. *Michael Ross and Barry Willer* 235

BOOK REVIEWS

Hospital Costs and Health Insurance, by Martin Feldstein. *Alain C. Enthoven* 242

Continuing Education in the Health Professions, by Robert Boissonnau. *Tricia C. Corbett* 246

Perspectives on Canadian Health and Social Services Policy: History and Emerging Trends, edited by Carl A. Meilicke and Janet L. Storch. *Eugene Vayda and Elaine Vayda* 248

Poverty and Health: Economic Causes and Consequences of Health Problems, by Harold S. Luft. *Richard M. Scheffler* 249

PUBLICATIONS RECEIVED 251

MEDICAL CARE
February 1982, Vol. XX, No. 2

Original Articles

Physician Performance in a Prepaid Health Plan:

Results of the Peer Review Program of the Health Insurance Plan of Greater New York

JEANNE MAGAGNA DEUSCHLE, M.A.,* BARBARA ALVAREZ, M.A.,† DONALD N. LOGSDON, M.D., M.P.H.,‡ WILLIAM M. STAHL, M.D.,§ AND HARRY SMITH, JR., PH.D.‖

A peer review program to evaluate the quality of medical care was established by the Health Insurance Plan of Greater New York in 1973. Physician performance is assessed through application of explicit process criteria to medical care as recorded in the patient record. A total of 6,788 records were reviewed in terms of the clinical management of acute otitis media, hypertension and breast lesions. Follow-up procedures to bring about positive changes in delivery of health care were integrated into the program at the time of the initial audit. Reaudit of 715 records was used to measure change in physician compliance with medical care standards. Results of this audit/reaudit process demonstrated statistically significant improvement in quality of clinical practice as measured by the assessment parameters.

NUMEROUS STUDIES have been conducted on the quality of health care in prepaid group practices and health maintenance organizations (HMOs). Cunningham and Williamson recently reviewed and evaluated literature on the quality of health care in an HMO as compared with that in other settings with similar populations.[1] The primary focus of these studies was the assessment of the quality of care. As expected, there were wide variations in the definitions of quality, the methodologies used and the criteria that were applied for the measurements. This article reports the changes demonstrated by a Peer Review Program audit and reaudit of ambulatory health care in a prepaid group practice HMO.

Background

The Health Insurance Plan of Greater New York (HIP) is one of the oldest and largest prepaid group practice health insurance plans in the country. In 1979, there were approximately 775,000 members receiving comprehensive health care on a prepaid basis through contractual arrangements that involved 26 multispecialty medical groups with 48 ambulatory centers in the metropolitan New York City area. HIP was approved as a New-York-State–certified HMO in 1979.

Numerous studies have been conducted at HIP on the quality of health care, includ-

* Director, Quality Assurance Division (QAD), Health Insurance Plan of Greater New York (HIP), New York, New York.

† Biostatistician, HIP, QAD.

‡ Director, Lifecycle Preventive Health Services Project, Sponsored by INSURE, New York, New York.

§ Professor of Surgery, New York Medical College, New York, New York.

‖ Chairman, Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York.

Reprint Requests: Jeanne M. Deuschle, Director Quality Assurance Division, Health Insurance Plan of Greater New York, 220 W. 58th St., New York, NY 10019.

0025-7079/82/0200/0127/$01.30 © J. B. Lippincott Co.

ing the early (1951) Makeover report,[2] the comparison studies by Shapiro on prematurity and perinatal mortality (1958 and 1960),[3,4] and patterns of medical care for an elderly, indigent population (1967).[5]

The current Peer Review Program at HIP was established in 1973 for the purpose of measuring the quality of medical care in each of the 28 (subsequently merged to 26) affiliated medical groups and to identify and resolve problems in the delivery of ambulatory health care by group physicians. The results reported in this paper are those of a large operational program with many practical methodologic constraints that were imposed because of the complex nature of the prepaid health system in which it was developed. The tracer methodology, as described by Kessner,[6] was utilized at HIP; the implementation of this methodology has already been reported.[7]

The first phase of the HIP evaluation process is based on the comparison of initial audit and reaudit results. This approach resulted in large-scale reviews (7,503 patient records) over a period of four years, involving multiple delivery sites—with proper sampling design, quality control for data collection, standardized analysis, presentation and follow-up of results.

Methods

Initial Audit Procedure

The HIP Peer Review Program assesses physician performance through application of explicit process criteria to medical care for selected tracer[6] conditions, as recorded in patient records.

For the first system assessment, medical care standards were established for hypertension in adults, acute otitis media in children and breast lesions. Criteria for each of these tracer conditions were formulated by a peer physician committee representing HIP and affiliated medical groups. Other areas for which medical care standards were subsequently developed are: urinary tract infections in adults, cervical neoplasia, anemia in children, acute pharyngitis in children, and follow-up of selected abnormal laboratory test values. The medical care standards are reviewed periodically for necessary modifications based on changes in medical management.

The basis for patient selection is the HIP Physician Visit Report Form. Historical information was obtained on the actual number of services provided by each of the 28 medical groups in the specialty areas involved, the total enrollment for each and estimate of the prevalence of the tracer condition. The size of the sample of patient's records for each medical group was determined by a simple random sampling formula, using an estimate of the standard deviation of quality scores within a group obtained from a pretest trial and establishing a 95 per cent confidence interval on the true group quality mean score of 10 quality points. If the prevalence figure for a group resulted in a small potential group population of patients for assessment, finite correction factors were used.

Data required for application of criteria are collected from patient records by professionally trained and supervised lay abstractors, using specially constructed data collection instruments.#

Following analysis (which utilized item importance weights[8] and is discussed in detail in another paper[7]), the Peer Review Program staff presents tables of results to the Peer Review Committee for discussion and formulation of recommendations. Medical groups are not identified in the presentations, except by code. This policy of confidentiality was established in an effort to ensure the Program's educational objectives. A Peer Review Committee report, which includes recommendations for

Careful analysis of the time required to complete the audits was made to ensure optimal utilization of manpower. The median time for review of a hypertension case was 32 minutes, for an acute otitis media case, 12 minutes, and for a breast lesion case, 20 minutes.

corrective action, item compliance and total median compliance, is then sent to the appropriate medical group director.

Members of the peer review staff, including a physician consultant, meet with the medical group's physician medical audit committee representatives for an in-depth discussion of the report and recommendations of the Peer Review Committee. In many instances, specific suggestions are made to assist the group in addressing identified problems.

The medical group director is required to respond in writing to the recommendations of the Peer Review Committee within two months. This response must include the medical group's plan to improve physician compliance with the medical care standards and, in particular, to address problem case listings that accompany the report. The plan of most medical groups included use of follow-up meetings with group physicians and/or conduct of mini-audits. In addition, changes in administrative procedures were sometimes used to facilitate resolution of problems.

Reaudit Procedure

The Peer Review Committee decided that a period of two months was adequate to permit medical groups to implement a physician education effort and to correct those problem areas identified in the reports and recommendations of the Peer Review Committee. In order to determine to what extent improvement had been made throughout the HIP system, a reaudit procedure was established, using the same medical care standards as for the initial audit.

Selection of Patient Records. Cases selected for reaudit were those diagnosed a minimum of two months after discussion of initial review results with medical group physicians.

Since it was desirable to assess improvement in physician compliance with medical care standards as soon as possible, it was decided to select a sample size large enough to estimate the true percentage of compliance within ± 5 percentage points, with a risk equal to 5 per cent at the system level (the 28 medical groups as an aggregate), rather than to estimate compliance for each of the medical groups. This resulted in a reaudit of a small number of cases (n = 13) from each medical group, with a total of n = 364 for the system, providing an adequate assessment of system improvement in physician compliance with medical care standards.

Data. For reaudit of acute otitis media, data were collected and analyzed in a manner identical to that of the original audit. However, for reaudit of hypertension, the data collection method was modified and was based on a combination of newly diagnosed and continuing care cases. Those newly diagnosed were collected and analyzed as in the original audit. The continuing care cases were selected randomly from those included in the initial audit. The collection and analysis of these cases were limited to assessment of compliance with standards of care for 1) longterm follow-up visits and 2) recording of blood pressure at each follow-up visit.

Of the 13 hypertension cases sampled in each medical group, nine were continuing care, and four were newly diagnosed. The smaller proportion of those newly diagnosed was decided upon because of the difficulty in identifying new cases of hypertension within a short period of time.

A systemwide reaudit for breast lesions was not indicated because of uniformly high physician compliance with medical care standards.

Reporting of Reaudit Results and Medical Group Follow-up. The reporting system described for presentation of initial audit data was also used for reaudit reporting, with minor modifications. For example, initial audit compliance was added to each table to facilitate comparison with reaudit and subsequent preparation of the Peer Review Committee report and recommendations. As was done for the initial

audit, peer review staff met with physician medical audit committee representatives of each medical group to discuss the reaudit report. For medical groups with continued "unsatisfactory" physician compliance with medical care standards, an additional group full-sample reaudit was conducted (as described for initial audit). The Peer Review Committee considered unsatisfactory performance as a median total compliance below 60 per cent on both audit and reaudit, as well as a statistically significant decrease in compliance on reaudit, regardless of the group's audit compliance.

Outcome analyses were considered in planning the scope of the peer review approach, but because of the complexity and limited strategies available to assess outcomes in operational settings, it was decided to limit initial efforts to review of the medical process. The collected process data do, however, include some intermediate outcome results. For example, for hypertension, blood pressure levels are recorded for all follow-up visits. Since we had not controlled for medical and

TABLE 1. Peer Review Program, Acute Otitis Media—26 Medical Groups, Comparison of Audit and Reaudit, Number and Per Cent of Cases in Which Physician Complied With Each Item Collected, Median Weighted Compliance by Category; Number of Cases in Initial Audit: 2106; Number of Cases Reaudited: 352

	Weight	Audit		Reaudit	
		Number	Per Cent	Number	Per Cent
Diagnostic Process	25%				
History					
Description of symptoms	5%	1422	67.5	289	82.3*
History of specific allergies	3%	793	37.7	205	58.4*
Physical					
Temperature	1%	581	27.6	157	44.6*
Tympanic membrane	5%	1226	58.2	292	83.2*
Cervical lymph nodes	3%	360	17.1	123	34.9*
Pharynx/tonsils	3%	1041	49.4	223	63.4*
Laboratory					
Culture and sensitivity	5%	17	24.6	7	41.2
Number of Applicable Cases†		(69)		(17)	
Therapeutic Process	75%				
Treatment					
Acceptable antimicrobial drug prescribed on diagnosis date	33.3%	1942	92.3	324	92.0
Adequate dosage recorded	16.7%	1166	61.5	230	72.6*
Follow-up					
Request return visit in 3–21 days	12.5%	1314	62.6	259	73.6*
Further evaluation	12.5%	62	73.8	8	53.3
Number of Applicable Cases‡		(84)		(15)	
Median Compliance					
Total		70.2		83.3*	
Diagnostic Process		50.4		70.0*	
Therapeutic Process		77.7		100.0*	
Grid Area 5—Median Total Compliance		50.0		66.9*	

* Statistically significant difference, $p < 0.05$.

† Applicable only if record of draining ear.

‡ Applicable if within the 12-month period prior to this episode child had three or more episodes or six or more visits for otitis media, or a recording of suspected hearing loss stated.

patient-related factors that also influence outcome, we have not presented these outcome results.

Results

Acute Otitis Media

Initial Audit Results. The 2,106 cases of acute otitis media (criteria for application of medical care standards are presented in Appendix A) reviewed in the initial audit of the HIP system were diagnosed between 1972 and 1977.

Table 1 presents the number and percentage of cases in which the physicians complied with each item and the median weighted compliance by category for the HIP system. In general, items in the diagnostic process, with 25 points of weight, were less frequently recorded than those in the therapeutic process, with 75 points of weight.

The median total compliance for the 2,106 cases was 70.2 per cent compliance for diagnostic process, 50.4 per cent, and for therapeutic process, 77.7 per cent. There was a great deal of variation in item compliance among the medical groups. For example, a recording concerning the status of the tympanic membrane, with 58.2 per cent overall compliance, ranged from 4–100 per cent among the individual medical groups. There was 92.3 per cent compliance for the item on acceptable antimicrobial drug prescribed on diagnosis date, with little variation among the medical groups. In most instances, noncompliance was attributed to no drug order being recorded in the patient record rather than to the prescription of an unacceptable drug. For 5.6 per cent, no drug was recorded on the date of diagnosis, and for only 1.4 per cent, the drug was unacceptable. Compliance with the drug item was not required if a physician impression of serous/viral otitis media was recorded.

There was 61.5 per cent compliance with the item "adequate dosage recorded." Noncompliance for this item was for the most part attributable to an absent/incomplete recording of dosage (34.1 per cent). For the remaining 4.3 per cent the dosage prescribed was judged to be below optimal levels.

A return visit in 3–21 days, with 62.6 per cent compliance, includes physician request for a revisit, regardless of whether the patient complied.

Figure 1 shows the median total compliance and the percentage of cases with total compliance of 50.0 per cent or higher for acute otitis media for the HIP system, as well as the distribution of the 28 medical groups for these parameters. The grid was originally developed as a management tool, so that improvement efforts could be focused on those medical groups falling into the lowest grid categories. This is discussed in detail in another article.[7]

The HIP system median compliance for acute otitis media was 70.2 per cent (Fig. 1). This tracer condition showed the greatest spread of the three for individual medical group compliance, which ranged for this condition from 33.5 per cent to 96.5 per cent.

Reaudit Results. The 352 cases of acute otitis media that were reaudited were diagnosed between 1976 and 1979.

Table 1 presents a comparison of audit and reaudit scores for individual items assessed, as well as the median compliance by category. Using the Mann-Whitney test,[9] statistically significant improvement ($p < 0.05$) was demonstrated for all median categories—total, increased 70.2–83.3 per cent; diagnostic process, 50.4–70.0 per cent; therapeutic process, 77.7–100.0 per cent.

Those items deemed most important to the medical management of the patient showed statistically significant improvement (binomial approximation to the normal, $p < 0.05$) or remained at the same high level of compliance as on initial audit, for example, prescribing an acceptable antimicrobial drug, 92.3 per cent on audit and 92.0 per cent on reaudit. Description of

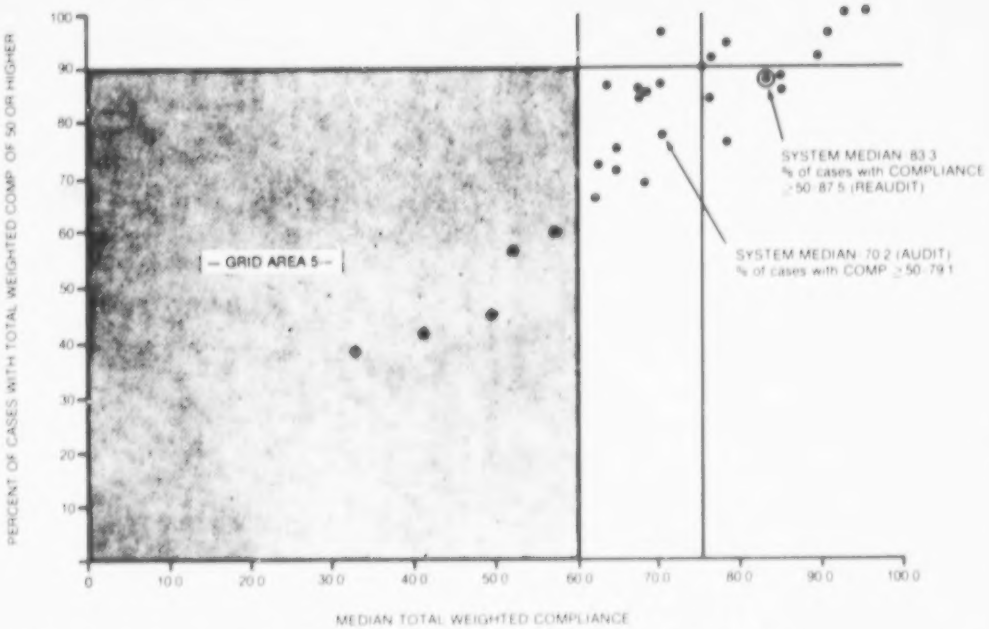


FIG. 1. Acute otitis media. Comparison of audit and reaudit (new cases); criteria for determining relationship of performance—28 HIP medical groups.

tympanic membrane on physical examination showed significant improvement—58.2 per cent on audit, and 83.2 per cent on reaudit. Compliance with the items on adequate drug dosage and a follow-up visit showed significant improvement. It is of interest that the noncompliance for adequate dosage was of the same kind as that found on initial audit, i.e., primarily absent or incomplete recording.

One item, further evaluation for those patients who had recurrent otitis media or hearing loss, decreased, but the result is not significant.

It is of particular importance that the median compliance for the aggregate of cases in the medical groups with audit compliance in the lowest grid category (see shaded area of Fig. 1) showed statistically significant improvement (Table 1)—50.0 per cent on audit to 66.9 per cent on reaudit (Mann-Whitney, $p < 0.05$).

Hypertension

Initial Audit Results. The 2,868 cases of hypertension (criteria for application of medical care standards are presented in Appendix A) reviewed in the initial audit of the HIP system were diagnosed between 1972 and 1977. Table 2 presents the number of cases in which the physicians complied with each item and the median compliance by category for the HIP system. Ninety per cent of the cases were classified as mild hypertension (diastolic 90–114 mm, systolic 140 mm or greater); 8 per cent as moderate hypertension (diastolic 115–130 mm); and 2 per cent severe hypertension (diastolic greater than 130 mm). In those instances in which more than one blood pressure was recorded at a visit, the reading with the lowest diastolic blood pressure was used. The median total compliance was 63.6 per cent; diagnostic process compliance was 55.0 per cent and

TABLE 2. Hypertension—26 Medical Groups, Comparison of Audit and Reaudit—New Cases: Number and Per Cent of Cases in Which Physician Complied With Each Item Collected
Median Weighted Compliance by Category; Number of Cases in Initial Audit: 2868*; Number of Cases Reaudited: 1114

	Weight	Audit		Reaudit	
		No.	%	No.	%
Diagnostic Process	40%				
Disease Detection	4%				
A blood pressure recorded within 2 years prior to diagnosis		1316	61.5	81	84.4‡
Number of Applicable Cases§		(2139)		(96)	
History	14.0				
Personal/social history	2.0	1903	56.2	83	75.5‡
Family history	4.0	1441	50.5	71	64.5‡
History of previous diagnosis/treatment of hypertension	4.0	1510	52.9	94	85.5‡
History of related symptoms	4.0	1483	52.0	74	67.3‡
Physical	7.0				
A blood pressure recorded within 2 months after diagnosis	1.0	1872	65.8	83	77.6‡
Weight	1.0	1625	56.9	79	72.5‡
Fundoscopic examination	2.0	581	20.4	25	24.0
Heart examination	1.0	2128	74.4	97	88.2‡
Abdominal description	1.0	1414	49.5	67	62.0‡
Extremities	1.0	1292	45.3	64	58.7‡
Laboratory and Other Tests	15.0				
Urinalysis	3.0	1888	66.0	83	77.6‡
Hematocrit or hemoglobin	1.0	1868	65.3	76	71.0
Blood urea nitrogen or serum creatinine	2.0	1597	55.9	76	74.0‡
Electrolytes—Serum potassium (prior to administration of antihypertensive drug therapy)	1.0	518	20.8	26	32.9‡
Blood sugar	1.0	1781	62.3	84	78.5‡
Serum cholesterol	1.0	1544	54.0	76	71.0‡
Electrocardiogram	3.0	1492	52.3	62	57.9
Chest x-ray	3.0	1742	61.0	81	74.3‡
Median Compliance					
Total		63.6		75.0‡	
Diagnostic Process		55.0		69.5‡	
Therapeutic Process		70.0		77.7‡	

	Weight	Audit		Reaudit	
		No.	%	No.	%
Therapeutic Process	60%				
Treatment					
Initial treatment regime		2147	74.9	98	88.3‡
Mild hypertension—5 week status	5.0				
Moderate hypertension—1 week status	6.0				
Severe hypertension—immediate status	30.0				
Six-month treatment/response status					
Mild hypertension only	5.0	378	60.4	17	63.0
Number of Applicable Cases		(626)		(27)	
Initial acceptable antihypertensive drug prescribed at acceptable dosage		1812	73.3	70	88.6‡
Mild hypertension	10.0				
Moderate hypertension	12.0				
Two-month response to drug therapy		1606	69.5	54	76.1
OR alteration of drug regime					
Mild hypertension	5.0				
Moderate hypertension	6.0				
Referral to consultant/ hospitalization/diastolic blood pressure reduced to 130 mm or less within 1 month of date of diagnosis					
Severe hypertension only	30.0	47	75.8	—	—
Number of Applicable Cases		(62)		(0)	
Serum potassium ordered after antihypertensive therapy initiated		226	9.3	7	10.0
Mild hypertension	2.5				
Moderate hypertension	3.0				
Potassium replacement ordered		73	21.4	1	— †
Mild hypertension	2.5				
Moderate hypertension	3.0				
Number of Applicable Cases #		(341)			
Follow-up (Mild and Moderate Only) Per Cent of Cases in Full Compliance					
Initial follow up	10.0	1272	44.8	57	51.8
Long-term follow up	10.0	1373	51.9	66	75.0‡
Blood pressure recorded at follow-up visits	10.0	2224	78.4	100	90.9‡
Grid Area 5—Median Total Compliance		58.8		76.6‡	

* In audit, 90.1% of cases reviewed were classified as mild hypertension, 7.6% as moderate, and 2.3% as severe.
† In reaudit, 93.7% of cases reviewed were classified as mild hypertension and 6.3% as moderate.
‡ Statistically significant difference, $p < 0.05$.
§ Applicable to patients enrolled in the medical group for at least 2 years prior to detection of hypertension.
- Percentage not reported, number of applicable cases less than 10.

that for therapeutic process was 70.0 per cent.

As was found with acute otitis media, there was wide variation in item compliance among the medical groups. For example, a recording concerning family history was found in 50.5 per cent of all cases, but the individual medical group compliance ranged from 21–89 per cent.

Compliance for funduscopic examination was 20.4 per cent, with a range of 1–38 per cent among the medical groups. This item, which was the most heavily weighted physical examination item in the diagnostic process, was the most poorly recorded. Ordering of serum potassium (20.8 per cent prior to administration of antihypertensive drug therapy) was another item low in physician compliance, ranging from 5–45 per cent among the individual medical groups. In general, there was good compliance with the initial therapeutic regimen¶ (74.9 per cent).

Physician compliance with the item concerning initial acceptable antihypertensive drug at acceptable dosage was also high (73.3 per cent), with a range of 56–90 per cent. Noncompliance with this item was as follows: antihypertensive drug or dosage not acceptable, 7.4 per cent; dosage incompletely recorded or illegible, 19.3 per cent. Compliance with the two-month response status to antihypertensive drug therapy was 69.5 per cent.**

There was full compliance with the initial follow-up standards†† in 44.8 per cent of the cases (i.e., all expected visits were requested by the physician), and 51.9 per cent of the cases were in full compliance with long-term follow-up standards.

Figure 2 shows that while there was some degree of spread among the medical groups, the majority clustered in the lowest grid areas. The range of medical group compliance was from 57.3–74.0 per cent.

Reaudit Results. The 111 new cases of hypertension that were reaudited were diagnosed between 1977 and 1979.

Table 2 presents a comparison of audit and reaudit compliance for individual items, as well as median compliance by category. Statistically significant improvement (Mann-Whitney, $p < 0.05$) was demonstrated for all median compliance categories: total, 63.6 per cent to 75.0 per cent; diagnostic process, 55.0 per cent to 69.5 per cent; therapeutic process, 70.0 per cent to 77.7 per cent.

As was shown for acute otitis media, those items deemed most important in terms of medical management of the patient either showed statistically significant improvement, using the normal approximation to the binomial, $p < 0.05$, or remained at the same high level of compliance found on initial audit. For example, in newly diagnosed cases of hypertension, both an acceptable initial treatment regime (74.9 per cent on audit to 88.3 per cent on reaudit) and acceptable antihypertensive drug/dosage (73.3 per cent on audit to 88.6 per cent on reaudit) showed siginificant improvement. Incomplete or nonrecording of drug/dosage was the principal reason for physician noncompliance with this item. There are two important items that, although showing some improvement, could improve further: 1) six-month re-

¶ *Initial treatment regime:* For cases classified as *mild* hypertension, any or all of the following were required with one month of index blood pressure reading: a follow-up visit, diet advice, sedation and antihypertensive drug treatment within one month of blood pressure reading. For cases classified as *moderate* hypertension, initiation of antihypertensive drug therapy was necessary within one week of index blood pressure reading. For cases classified as *severe* hypertension,immediate initiation of antihypertensive drug therapy or referral to consultant or hospitalization was required.

** *Two month response status* required 1) the diastolic blood pressure to be reduced by 10 mm; if only systolic hypertension, the systolic blood pressure to be reduced by 10 mm unless physician stated that such reduction was contraindicated; and 2) if no response by blood pressure reduction, then drug dosage to be modified.

†† *Follow-up standards required: Initial follow up*–at least one visit every two months for four months for cases classified as mild hypertension and one visit every month for three months for cases classified as moderate; *Long-term follow up*–one visit every six months for both mild and moderate hypertension.

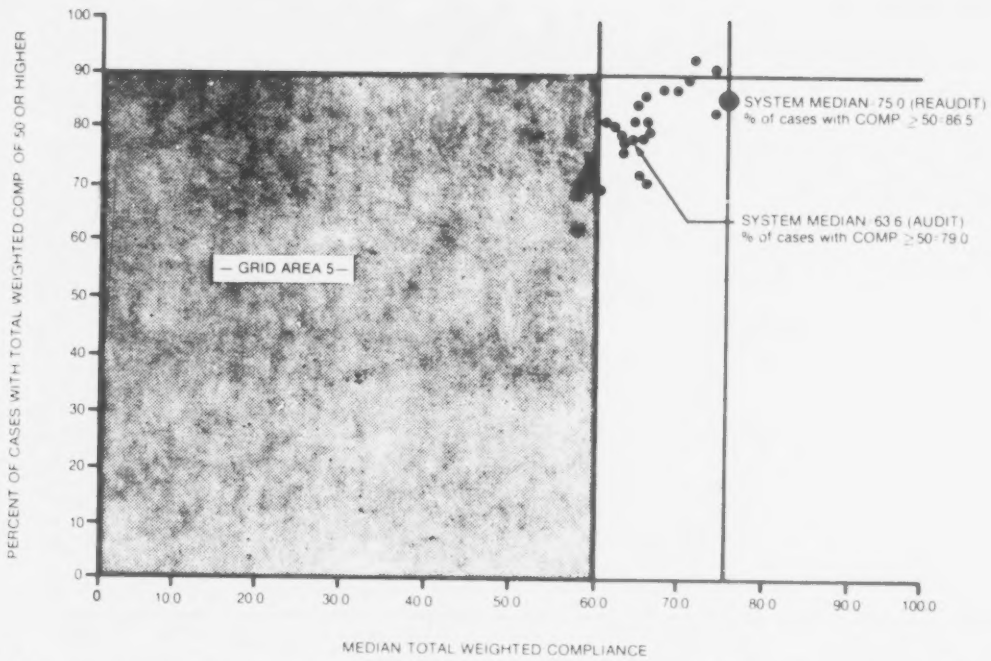


FIG. 2. Hypertension. Comparison of audit and reaudit (new cases); criteria for determining relationship of performance—28 HIP medical groups.

sponse to treatment for patients with mild hypertension who were not initially placed on an antihypertensive drug regime [six-month response status (for patients with mild hypertension who were not initially placed on an antihypertensive drug regime) required reduction of blood pressure to normal range within six months of initiation of antihypertensive drug therapy] and 2) full compliance with standards for initial follow-up visits.

Of the 19 items included in the diagnostic process, 16 showed statistically significant improvement. Physician compliance with medical standards for funduscopic examination and pre- and posttreatment electrolytes remained exceptionally low.

As was found for acute otitis media, the median total compliance for the aggregate of cases for those medical groups with audit compliance in the lowest grid category (see Fig. 2, shaded area) showed statistically significant improvement (Table 2)–58.8 per cent on audit to 76.6 per cent on reaudit (Mann-Whitney $p < 0.05$).

A sample of cases from those included in the initial audit was also reaudited for hypertension in order to assess the status of continuing care (Table 3). Two items were reaudited: 1) physician compliance with medical care standards for long-term follow-up where improvement was measured in terms of full compliance with all expected follow-up visits and 2) recording of blood pressure at all visits. Two hundred fifty-two cases included in the initial audit were reaudited. Statistically significant improvement (McNemar test,[9] $p < 0.05$) in physician compliance with medical standards was demonstrated for both long-term follow-up (48.1 per cent on audit to 63.6 per cent on reaudit) and recording of blood pressure at each visit (77.9 per cent on audit to 90.6 per cent on reaudit).

TABLE 3. Peer Review Program, Hypertension—26 Medical Groups, Comparison of Audit and Reaudit—Continuing Care (n = 252 Cases)

	Audit		Reaudit	
	No.	%	No.	%
Cases in Full Compliance				
Long-term follow-up	111	48.1	147	63.6*
Number of Applicable Cases†	(231)		(231)	
Blood pressure recorded at visits	166	77.9	193	90.6*
Number of Applicable Cases†	(213)		(213)	

* Statistically significant difference, $p < 0.05$.

† The method used derived 252 cases, of which 231 had sufficient time (i.e., 6 months) to assess compliance with at least one long-term follow-up visit. Of these 231 cases, only 213 had at least one visit, thereby allowing a blood pressure recording to be assessed.

Breast Lesions

Initial Audit Results. The 1,814 cases of breast lesions reviewed in the initial audit of the HIP system were diagnosed between 1968 and 1977.

The cases were classified as either benign or suspicious, based on the physician's statement describing the lesion and the diagnostic impression as recorded in the patient's record. Only cases with a diagnostic impression of benign breast disease, e.g., fibrocystic disease, were classified as benign. All others, e.g., mass, lump, were classified as suspicious. A surgeon's diagnosis was given priority over a diagnosis by a family physician. In the occasional instances for which it was not possible for the lay abstractor to classify a case, it was reviewed by the Program's physician consultant. In 1,108 cases (61.1 per cent), the lesions were classified as benign, and in 706 cases (38.9 per cent), they were classified as suspicious (Table 4).

Physician compliance for breast lesions was much higher than for acute otitis media and hypertension. The median total compliance was 87.5 per cent. Compliance for the diagnostic process was 73.3 per cent and for the therapeutic process, compliance was 100.0 per cent.

For the diagnostic process, physician compliance in recording the history items was low, with essentially little difference among the medical groups. The three physical examination items assigned a total weight of 17 points were in general well performed, in particular, description of the breast lesion (89.6 per cent), which required for compliance that the physician record the size, location and right/left breast. Physician compliance with the ordering of mammography was high (88.4 per cent). This item was applicable only to those women who were in follow-up for the breast lesion for a minimum of two years.

Physician compliance with the therapeutic process standards for breast lesions classified as benign required only follow-up visits (initially every three months for six months, then every six months), with a description of the lesion at each visit. There were 500 cases (45.1 per cent) in full compliance with this standard. This means that all expected visits, as well as a description of the lesion, were recorded.

Physician compliance with medical care standards for lesions classified as suspicious required a biopsy within two months. If a biopsy was performed between two and six months, partial compliance was assigned. There was very high physician compliance with this standard—597 cases (84.6 per cent).

Of the 703 cases classified as "suspicious" (three cases deleted from suspicious

TABLE 4. Peer Review Program, Breast Lesion Results—28 Medical Groups—Initial Audit, Number and Per Cent of Cases in Which Physician Complied With Each Item Collected; Median Weighted Compliance by Category
(Number of Cases Reviewed: 1814)

	Weight	Number	Per Cent
Diagnostic Process	30%		
Disease Detection			
Breast examination within 2 years prior to diagnosis of initial breast lesion	6%	742	59.7
Number of Applicable Cases*†		(1242)	
History			
Previous breast disease or breast surgery	1%	120	6.6
Hormone therapy	1%	185	10.2
Menstrual history	1%	418	23.0
Physical			
Examination of both breasts	4%	1291	71.2
Regional lymph nodes	3%	862	47.5
Description of breast lesion (size, location, and right/left breast)	10%	1626	89.6
Other			
Mammography	4%	1017	88.4
Number of Applicable Cases*¶		(1151)	

	Weight Assigned to Each Type	Number and Per Cent in Each Category	
		Number	Per Cent
Therapeutic Process			
Total Weight	70%		
Benign lesion‡ (No biopsy required)	0–70%		
Number and per cent of cases in *Full*§ (100%) compliance with follow-up standards		500	45.1
Suspicious lesion‡ (Biopsy required)	0–70%		
Number and per cent of cases in *Full*§ (100%) compliance with therapeutic process standards		597	84.6
Median Compliance	Weight		
Total	100%	87.5	
Diagnostic Process	30%	73.3	
Therapeutic Process	70%	100.0	

* The number of applicable cases is only listed for items on which the denominator of the per cent differs markedly from the total number reviewed.

† Applicable to those patients who were over 35 years of age and enrolled in the medical group for at least two years prior to detection of breast lesion.

‡ Number and per cent of lesions classified as: benign lesions: n = 1108; % = 61.1;
suspicious lesions: n = 706; % = 38.9.

§ Full compliance for benign lesions = follow-up visits with complete description of lesion at frequency specified in standards.
Full compliance for suspicious lesions = as for benign and biopsy within two months.

¶ Applicable if performed or to those cases that had two full years of review available.

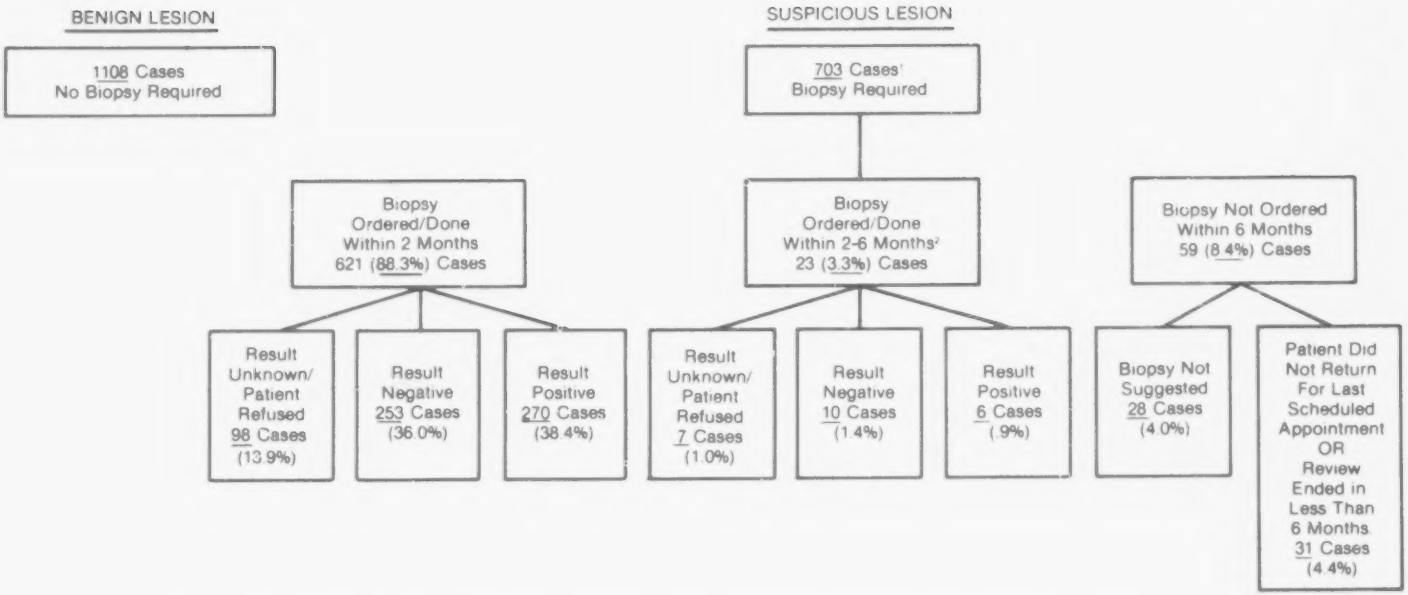


FIG. 3. Peer Review Program. Flow chart of results of breast lesion cases reviewed—HIP system (n = 1,814).

[1] There were three additional cases classified initially as suspicious by a surgeon. Lesion not present at subsequent examination by surgeon.

[2] In a few instances, the time period may actually be up to two years if last visit in an interval is a broken appointment.

classification—see Fig. 3), biopsy was suggested/ordered within two months for 621 cases (88.3 per cent) and within two to six months for 23 (3.3 per cent) (Fig. 3).

Of the three tracer conditions reviewed, breast lesions showed the least spread in compliance, with all but one medical group in the highest grid category (Fig. 4).

Reliability of Audit and Reaudit Data

Procedures were implemented in the ongoing system to measure and assure the reliability and accuracy of the data collected. Following these procedures, the field supervisor, on a daily basis, selected a random sample (approximately 5 per cent) of the patient records that had already been abstracted and resubmitted them to be read a second time by another abstractor. In addition, a subsample of these duplicate records was selected periodically and read a third time by the field supervisor.

Data obtained from both audit and reaudit resulted in a reliability rate for abstracted items exceeding 90 per cent, consistent with the reproducibility rate for lay abstractors of medical records reported by Lyons and Payne[8, 10] and Hulka.[11]

Discussion

As Brook has pointed out, "No new conceptual frameworks for assessing quality have been developed during the last two decades." Assessment of structure, process and outcome variables are all used but, as yet, ". . . there is no evidence to indicate which one produces the most valid judgments of the program or practice being assessed . . . opinion differs as to which type of judgment is the most valid."[12]

Largely because of its feasibility in terms of ". . . objectivity, economy of data collection, and adaptability to computer analysis, as compared to assessment of outcomes or the use of physician reviewers,"[13] HIP's Peer Review Program implemented the most commonly used method to assess physician performance—the use of explicit process criteria for selected conditions.

Criteria or standards developed by peer consensus are compared with information available from the patient record. Christoffel has discussed limitations to this approach in detail.[14] A frequently cited limitation is concerned with problems encountered in the use of the medical record, primarily lack of documentation of medical management; examples include the common practice of nonrecording of "negatives" or normal findings, failure to record drugs and dosages, and failure to record requests for return visits—all of which can bias results to noncompliance.

In spite of the aforementioned problems, the authors are in agreement with Payne's postulate, ". . . that use of the medical record for measuring physician performance will improve communication, improve the quality of medical care, and improve the content of the medical record."[15] A positive relationship between good recording and medical care performance has already been demonstrated by Lyons and Payne.[16]

Related to this issue is the frequent comment that this type of peer review activity—application of explicit process criteria to the medical record—primarily addresses *recording* of medical care, and that improvement on reaudit indicates improved recording. Although our results do indicate that a significant amount of noncompliance is attributable to nonrecording, there are a number of examples of *recorded management* that were not in compliance with the standards. Several examples showing the distribution of noncompliance follow.

Acute Otitis Media—System Audit

Noncompliance with the item "acceptable antimicrobial drug prescribed on diagnosis date" was 7.7 per cent. Presumably, most of this noncompliance was related to managerial rather than to recording prob-

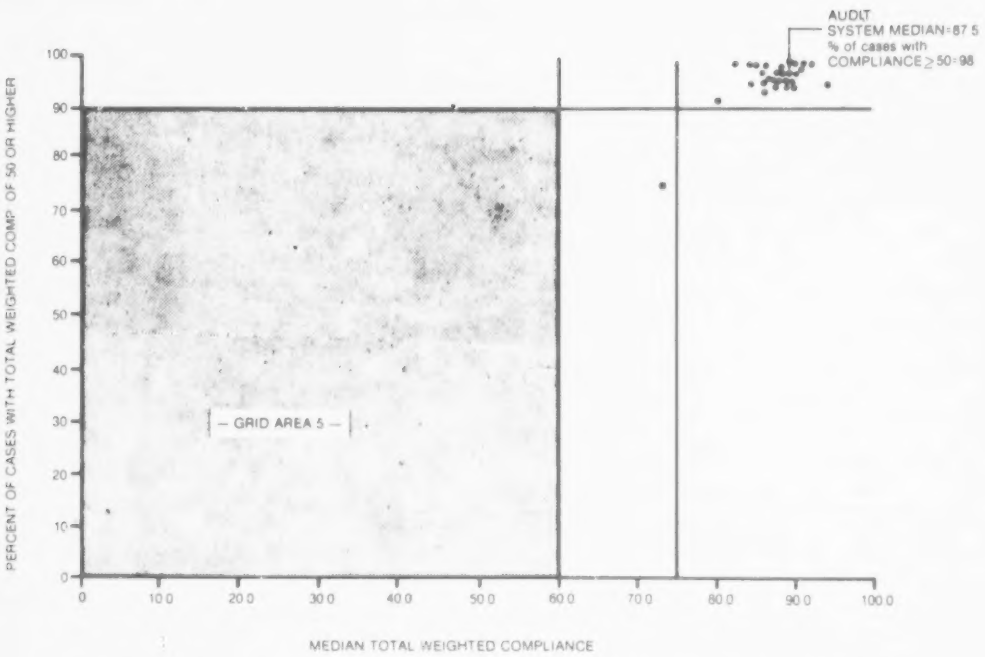


FIG. 4. Breast lesions. Criteria for determining relationship of performance—28 HIP medical groups.

lems, as follows: 5.6 per cent, no antimicrobial drug prescribed; 1.4 per cent, unacceptable antimicrobial drug prescribed (usually a cephalosporin; subsequent modification of the standards considered these drugs acceptable although not recommended); 0.6 per cent, penicillin prescribed in spite of physician's acknowledged statement of child's penicillin allergy. Noncompliance with this item remained unchanged on reaudit.

Noncompliance with the item "adequate dosage prescribed" was 38.4 per cent. While 34.1 per cent of this noncompliance was attributable to nonrecording of dosage, 4.3 per cent was because the dosage prescribed was judged to be below optional levels. Noncompliance with this item on reaudit decreased significantly, from 38.4 per cent to 27.5 per cent. The distribution of noncompliance showed impact on recording only.

Hypertension—System Audit

Noncompliance with the item "initial acceptable antihypertensive drug/dosage" was 26.7 per cent. While 19.3 per cent of this was attributable to nonrecording of drug dosage, 7.4 per cent was because of prescribing unacceptable drug or dosage. Essentially all of this was unacceptable dosage—only 0.5 per cent was unacceptable drug. It should be mentioned that dosage criteria did not accept a dosage prescribed for fewer than five days per week. Dosage criteria were made discretionary in a 1979 modification of these standards. Noncompliance with this item decreased significantly on reaudit, from 26.7 per cent to 11.4 per cent. The distribution of noncompliance showed impact both on recording and on prescribing of acceptable drug/dosage.

* * *

Figures 1, 2 and 4 show a significant difference in system compliance for the three conditions—acute otitis media, hypertension and breast lesions. This result implies that the system quality of medical care varies by condition, with breast lesions having the highest system compliance. It can be postulated that the high physician compliance with medical care standards for breast lesions may be related to characteristics of the condition (a malignant breast lesion produces greater morbidity than does hypertension or acute otitis media and, therefore, may be more carefully managed) or to characteristics of physicians in specialty practice (medicine and pediatrics versus surgery). Other factors that may also contribute to these differences in condition compliance are the number and kinds of items in each of the standards and the item weights assigned.

Brook has pointed out that "the most serious threat to the positive effect of quality assurance programs on care and health status lies with failure to use the results of quality assessment studies to improve care. . . . Effective quality assurance programs must take full advantage of feedback mechanisms, such as the reaudit procedures . . . to ensure quality and improve health status."[12] From its inception, the HIP Peer Review Program has been aware of the importance of building into the assessment of physician performance a specific approach to assist medical group physicians in resolving identified problems. The results of audit/reaudit presented in this paper have demonstrated that a significant increase in physician compliance with the medical care standards was obtained following use of a feedback mechanism.

We fully recognize that causal statements cannot be made concerning the impact of the Peer Review Program on this improvement. Organizational constraints did not permit the development of the sophisticated research design that would be required to address this issue. However, HIP is encouraged by these results, and it is our impression that the activity of the Program provides a significant impetus for improvement in physician compliance with medical care standards.

In line with the continued need to find ways to use results for the improvement of medical care, the Peer Review Program is exploring mechanisms to encourage and assist the groups in the development of group-based quality assurance activities that can work in a complementary way with HIP's systemwide quality assurance program. It is our opinion that this change in quality assurance emphasis, from a system responsibility to a shared local responsibility with the medical groups, will be a catalyst to improved quality of medical care in the HIP system.

NOTE: Medical care standards and data collection methods are available from Mrs. Deuschle upon request.

References

1. Cunningham FC, Williamson JW. How does the quality of health care in HMOs compare to that in other settings? An analytic literature review: 1958 to 1979. The Group Health Journal 1980;1.

2. Makover HB. The quality of medical care, methodology of survey of the medical groups associated with the Health Insurance Plan of New York. Am J Public Health 1951;41:824.

3. Shapiro S, Weiner L, Densen PM. Comparison of prematurity and perinatal mortality in a general population and in the population of a prepaid group practice, medical care plan. Am J Public Health 1958;48:170.

4. Shapiro S, Jacobziner H, Densen PM, Weiner L. Further observations on prematurity and perinatal mortality in a general population and in the population of a prepaid group practice medical care plan. Am J Public Health 1960;50:1304.

5. Shapiro S, Williams JJ, Yerby AS, Densen PM. Patterns of medical use by the indigent aged under two systems of medical care. Am J Public Health 1967;57:784.

6. Kessner DM, Kalk CE. A Strategy for Evaluating Health Services. Contrasts in Health Status. Washington, D.C.: National Academy of Sciences, Institute of Medicine, 1973;2.

7. Deuschle JM, et al. Implementation of a peer review system for ambulatory care: Experience of the Health Insurance Plan of Greater New York. Public Health Rep 1978;93:258.

8. Lyons TF, Payne BC. The use of item importance weights in assessing physician performance with predetermined criteria indices. Med Care 1975;13:432.

9. Conover WJ. Practical Nonparametric Statistics. New York: John Wiley & Sons, 1971.

10. Payne BC, Lyons TF, et al. The Quality of Medical Care: Evaluation and Improvement. Health Services Monograph Series. Chicago, Ill.: Hospital Research and Educational Trust, 1976.

11. Hulka BS, et al. Peer review in ambulatory care: Use of explicit criteria and implicit judgments. Med Care 1979;17(suppl).

12. Brook RH, Davies-Avery A. Mechanisms for Assuring Quality of U.S. Medical Care Services: Past, Present, and Future. Prepared under a grant from Department of Health, Education, and Welfare, R-1939-HEW. Santa Monica: The Rand Corporation, 1977.

13. Rosenthal G. Foreward to NCHSR Research Summary Series. Quality of Ambulatory Care: An Exploration of The Discrepancy Between Expiicit Process Criteria and Performance (Hulka). U.S. Department of Health, Education, and Welfare, Public Health Service, Office of Health Research, Statistics, and Technology, National Center for Health Services Research, 1980. [DHEW Publication No. (PHS) 80-3244.]

14. Christoffel T, Loewenthal M. Evaluating the quality of ambulatory health care: A review of emerging methods. Med Care 1977;15:877.

15. Payne BC. The medical record as a basis for assessing physician competence. Ann Intern Med 1979;91:623.

16. Lyons TF, Payne BC. The relationship of physicians' medical recording performance to their medical care performance. Med Care 1974;12:463.

Appendix A

1. Acute Otitis Media

The criteria used for application of the medical care standards were physician diagnosis of acute otitis media or, in the absence of physician diagnosis of acute otitis media on a patient's record, recording of painful ear, inflammation/erythema of tympanic membrane, diminished light reflex, fluid in middle ear cavity or bleb on tympanic membrane.

2. Hypertension

The criteria used for application of the medical care standards were a blood pressure reading at the following age and blood pressure levels:

Age	Systolic and/or Diastolic up to or Exceeding
18–44	140/90mm
45–64	150/95mm
65 and over	160/95mm

MEDICAL CARE
February 1982, Vol. XX, No. 2

Episode-Based Efficiency Comparisons for Physicians and Nurse Practitioners

DAVID S. SALKEVER, PH.D., ELIZABETH A. SKINNER, M.S.W., DONALD M. STEINWACHS, PH.D., AND HARVEY KATZ, M.D.

Most previous studies comparing the efficiency of new health practitioners with that of physicians have used the visit as the basic unit of output. Several researchers have noted, though, that the episode is a conceptually superior output unit in several respects, although it is more complex to deal with methodologically. This study demonstrates the application of episode-based methods for comparing the efficiency of physicians with that of nurse practitioners. Data are drawn from the information system of the Columbia Medical Plan and from observations of provider time inputs. The analysis is confined to care episodes for otitis media and sore throat in the Department of Pediatrics. Results indicate that per episode costs with nurse practitioners as the initial provider are approximately 20 per cent below the costs of episodes in which physicians are the initial provider. Examination of a limited amount of data on patient-reported measures of effectiveness indicates that while nurse practitioners' care is less costly, it is not less effective. These findings are particularly interesting in light of recent doubts expressed about cost-savings from using new health practitioners, and particularly nurse practitioners, in group practice settings.

A VARIETY OF APPROACHES have been used recently to compare the costs or productivity of physicians with that of new health practitioners (NHPs) in providing primary care. The most important difference among these approaches concerns the way in which the unit of output is defined. A number of studies have defined output in terms of a collection of specific types of tasks or visits. For example, Record et al.[1] divided all outpatient visits to the Kaiser-Portland internal medicine department into 23 categories and defined annual department outpatient output as the annual frequencies for each of these 23 types of visits. Costs of producing this output with physicians alone and with the least-cost combination of physicians and physician's assistants were then compared. An earlier study of primary care practices by Smith et al.[2] used 41 categories of visits to define output and examined the increase in output (i.e., the proportionate increase in visits of all types) that could be obtained by adding a physician's assistant to a practice.

Other studies using the visit as the basic unit of output have examined only one or several particular types of visits. Examples are the recent comparisons of well-baby care provided by pediatricians and pediatric nurse practitioners.[3–8] Still other analyses based on visits have not differentiated among visits by type or content of

From the Health Services Research and Development Center, The Johns Hopkins Medical Institutions, Baltimore, Maryland.

This research was supported by Grants HS00427 and HS01964 from the National Center for Health Services Research.

Address for Reprints: Elizabeth A. Skinner, M.S.W., Health Services R and D Center, The Johns Hopkins Medical Institutions, 624 N. Broadway, Baltimore, MD 21205.

0025-7079/82/0200/0143/$01.05 © J. B. Lippincott Co.

services. This is most common in before–after comparisons of cost per visit for practices that have added NHPs, or similar comparisons between practices with and without NHPs.[7,8] However, the productivity or efficiency implications of these comparisons are clouded by the possibility that the mix of visits is altered when NHPs are introduced into a practice.[9,10] If this mix changes, before–after or interpractice differences in cost per visit may be partly due to differences in the mix of visit types rather than to productivity differentials.

An alternative unit of output less frequently employed than the visit is the person-year of care (i.e., care of one person for a one-year period). This unit is only appropriate for situations in which defined comparison groups of patients are assigned either to a physician or to an NHP, who serves as the primary source of first-contact care and who initiates the referral process to other providers. Average cost per person per year for all services (or a defined subset of services) in each comparison group is used as the efficiency criterion. Examples are the evaluation by Chambers et al. of family practice services in rural Newfoundland[11] and the studies of chronically ill patients reported by Spector et al.[12] and Lewis and Resnik.[13] The most appealing feature of this type of cost comparison is that it is population-based. However, as already noted, it can only be applied in certain rather unusual circumstances. Moreover, such cost comparisons will clearly be sensitive to random variations in the incidence of health problems among comparison groups, so that even in a randomized trial results may be unreliable unless the comparison groups are quite large.

Another unit of output that has several conceptual advantages for cost and productivity comparisons is the episode of care. It has been pointed out that episode-based comparisons are consistent with the idea that treatment of health problems, rather than the provision of individual services, is the primary objective of medical care.[14,15] From this perspective, the episode is preferable to the visit as the unit of comparison because it takes into account differences in referral and revisit rates between physicians and NHPs. The episode is also a more appropriate unit for measuring differences in effectiveness of care, since the outcome of the care process may be causally related not only to service received at a single visit, but to any services received over the course of the episode. However, episode-based comparisons are methodologically more complex than comparisons based on visits, since procedures must be developed for defining episodes and identifying the services contained in an episode. Perhaps because of this added complexity, and in spite of their conceptual advantages, episode-based cost or productivity comparisons of physicians and NHPs have not been carried out previously. (A recent study by Moscovice does report episode-based comparisons between nurses and family nurse practitioners in treating otitis media and urinary tract infection.[16]) Research by Tompkins et al.[17] compared the cost-effectiveness of internists and algorithm-assisted physician's assistants in treating adult respiratory illnesses using a 10–14-day period as an approximation for the episode of care.

In this article we shall describe the methods and findings of efficiency comparisons between pediatricians and pediatric nurse practitioners in a group-practice HMO in treating episodes of otitis media and sore throat. A major objective of this research was to develop and demonstrate practical procedures for episode-based comparisons that maximize the use of routinely collected data and minimize interruption of patient and provider activities. It is suggested that such procedures make it feasible to carry out episode-based comparisons for a variety of types of health problems and in a variety of organized ambulatory care settings. The

research reported here is also of substantive interest since the extent of cost or productivity gains from NHP employment in group practices is currently in doubt, with some analysts reporting large gains and others suggesting a much smaller impact.[18] Furthermore, it has been argued that existing evidence of cost or productivity gains for nurse practitioners is weaker than that for other types of NHPs.[18]

Study Setting

The site for the study was the Department of Pediatrics of the Columbia Medical Plan (CMP), an HMO serving approximately 23,000 enrollees in the city of Columbia, Maryland. The Department of Pediatrics was staffed by four "full-time equivalent" (FTE) physicians, four FTE nurse practitioners and approximately four health assistants. The department has employed nurse practitioners and health assistants since late 1969. The annual volume of patient visits is now approximately 24,000, with physicians being the primary provider in about 60 per cent of these visits and nurse practitioners being the primary provider in about 40 per cent of the visits.

Nurse practitioners and physicians both function as independent professionals in providing ambulatory care. Health assistants may perform a few well-defined specific clinical tasks (such as immunizations), but their primary role is in scheduling, telephone answering and clerical functions. They also carry out preliminary tasks during physician visits, such as bringing the patient and the patient's record to the examination room, weighing and temperature taking. At the time of this study, nurse practitioners usually performed these tasks for their own patients.

CMP's information system is composed of two parts. The enrollment file contains data on enrollees' demographic characteristics, while the encounter system maintains information on the nature and content of each visit. Data are entered from encounter forms completed at each visit and include patient name and enrollment number, appointment status, purpose of the visit, providers seen, laboratory, x-ray and other procedures ordered, problems or diagnoses, and disposition. Prescriptions filled by the plan's in-house pharmacy are entered as well. It is estimated that all but 5 per cent of prescriptions written by CMP physicians are captured by the system.[19]

Research Methods

As noted earlier, a primary objective of the research was to develop a relatively inexpensive procedure for estimating costs of episodes of care. To achieve this objective, our study plan incorporated findings from previous methodological research on the definition of episodes and relied heavily on the routinely collected data in the CMP information system. However, one important element in the cost estimation of ambulatory care episodes, which was not captured by the information system, was provider time inputs. (Duration of visit is recorded on the plan's encounter forms, but only in 15-minute intervals.) After consideration of alternative data collection methods, it was decided that the most economical and nonintrusive procedure for determining provider time inputs was to employ observers stationed outside of offices and treatment rooms who would record the comings and goings of providers and patients. This procedure was followed during a two-week period in May of 1976. In each observation period, observers were assigned to particular providers and recorded data on all visits to these providers. The observation schedule was designed to provide data on both walk-in and scheduled visits to each of the physicians and nurse practitioners in the Department of Pediatrics. Roughly half of all visits to the department during the two-week period were captured by the observers. (A

more detailed description of the methods used in the observational study has been reported elsewhere.[20])

While the research focused on comparisons of episode costs, information was also available, from prior research at the CMP, on the quality and effectiveness of care provided by physicians and nurse practitioners. This was gathered as part of an Experimental Medical Care Review Organization (EMCRO) project. Data collection methods and analyses developed from this project have been reported previously.[21,22] In our research, further data analyses dealing specifically with otitis media and sore throat episodes were carried out.*

Definition and Selection of Episodes

Episodes of care for otitis media and sore throat were identified using the encounter data information system. Criteria were applied to link individual visits into episodes using diagnostic information, time interval, disposition and type of visit (initial illness, follow-up illness, etc.). A maximum interval of three weeks was allowed between successive visits for sore throat and purulent otitis media, while five weeks was used between serous otitis media visits. The selection of these intervals was based on clinical consensus, with comparisons between the information system and medical record to test the methodology.[23] The results indicate that descriptions of episodes of care derived from the information system using explicit criteria and those obtained through review of the medical chart are comparable. Exceptions are those cases in which otitis media is chronic and the distinction between individual episodes in not clear either in the chart or the information system. It should also be noted that episodes may involve more than one condition (e.g., otitis media and upper respiratory infection), which may be a significant determinant of longer episodes and the more intensive use of resources.

The episode definitions were applied to encounter system data for all visits to the Department of Pediatrics during the period of January through June, 1976. As a result, 409 otitis media episodes involving 646 visits, and 390 sore throat episodes involving 409 visits were identified.† These episodes did include a total of 17 visits to the CMP's Urgent Care Center, a walk-in clinic for after-hours services in which almost all care was rendered by physicians. However, no episodes had more than one Urgent Care visit, and all episodes where the initial visit was to Urgent Care were excluded from the study. Although it is possible that some visits to ENT specialists involved treatment of the two problems under study, it should be noted that linking these visits with specific episodes is problematic because ENT referrals usually are occasioned by repeated episodes or chronic problems. Using our method for defining episodes, only 20 ENT visits were identified in the 799 study episodes. Visits to other departments in the plan were presumably not related to otitis media or sore throat problems, and were not included in the episodes.

Time Inputs and Associated Costs

The costs of an episode of care were broken down into two components: 1) costs associated with provider time inputs and 2) costs of ancillary services and drugs. Costs associated with provider time inputs for

* While the EMCRO data on pediatric patients at the CMP were gathered two years prior to the other data used in our study, available evidence suggests that the characteristics of the patient population and practice patterns were essentially unchanged over this period.

† Note that episodes apparently beginning in early January or extending to late June could not be clearly delimited with the January–June data, based on our decision rules, they might in fact have begun at an earlier date or ended at a later date. Therefore, these episodes were excluded from the study.

TABLE 1. Mean Patient Contact Time Per Visit

	MD		NP	
	Time (min.)	N	Time (min.)	N
Otitis media				
Initial visits*	9.3	20	13.4	14
Follow-up visits	7.5	18	8.4	5†
Sore throat				
Initial visits‡	7.7	16	13.1	18
Follow-up visits	11.0	3	9.5	2

* MD–NP difference is significant at the 0.10 level (two-tailed).
† Excludes one six-minute visit with MD consultation.
‡ MD–NP difference is significant at the 0.05 level (two-tailed).

the two types of episodes studied here were calculated on a per-visit basis for six categories of visits: initial and follow-up visits to physicians and to nurse practitioners with and without physician consultation. The first step in this calculation was to compute the average provider time spent in direct patient contact for each type of visit. These figures were based on the May 1976 observational study. From a total of 379 visits for which time data were obtained, 58 were for treatment of otitis media and 39 were for treatment of sore throat. The mean times for initial and follow-up visits by provider type are shown in Table 1. For both problem categories, physician initial visit times were considerably shorter, while physician and nurse practitioner times for follow-up visits in the data were similar. However, the numbers of follow-up visits in the data were very small, and only a single nurse practitioner visit with physician consultation (a follow-up otitis media visit) was recorded. Therefore, the cost calculations for nurse practitioner visits with physician consultations used the nurse practitioner times shown in Table 1 and assumed a physician contact time of 4.6 minutes. (The 4.6 minute figure was the average physician contact time for all nurse practitioner visits with physician consultation, regardless of problem type, recorded in the observational study.) Furthermore, because of the limited numbers of otitis media and sore throat visits observed, the sensitivity of study findings of changes in mean time values was also examined and is discussed below.

Noncontact provider time inputs for each category of visit were assumed to be proportional to direct contact time. The proportionality factors used were the ratio of noncontact time to contact time during walk-in hours observed in the two-week study. Data for walk-in hours rather than scheduled-visit hours were used because 90 of the 97 observed otitis media and sore throat visits occurred during walk-in hours. These ratios were 0.84 for physicians and 0.73 for nurse practitioners.

For visits including physician time inputs, the services of health assistants must also be considered. Since detailed data on health assistants were not obtained in the observational study, we assigned 0.5 minutes of health assistants' time for each minute of physician time. The 0.5 figure was based on the fact that one health assistant was assigned to assist two physicians in their patient care activities, while other health assistants carried out scheduling, telephone and clerical duties.

The final component of costs associated with provider time inputs is space costs. Each pediatrician utilized 280 square feet

of space in one office and two examining rooms, while each nurse practitioner was assigned an office with 120 square feet of space. Assuming that space was utilized for 2,000 hours annually and employing an annual cost figure (supplied by the plan's financial office) of $9.50 per square foot, space costs per minute of physician and nurse practitioner time were figured at 2.22 cents and 0.95 cents, respectively.

In summary, the formula for provider-time-related costs at each visit is:

$$(T_{MD} + 0.84T_{MD})\ (W_{MD} + 0.5W_{HA} + 2.22) + (T_{NP} + 0.73T_{NP})\ (W_{NP} + 0.95),$$

where T_{MD} and T_{NP} are the minutes of the physician and nurse practitioner direct contact time, and where W_{MD}, W_{HA}, and W_{NP} are the salary and fringe benefit costs per minute for physicians, health assistants and nurse practitioners. The plan's financial data indicated values of 40 cents for W_{MD}, 8 cents for W_{HA} and 15 cents for W_{NP}. Inserting these values and the T_{MD} and T_{NP} figures reported into this formula yielded the cost figures shown in Table 2. (Note that $T_{MD} = 0$ for nurse practitioner visits without consultation and $T_{NP} = 0$ for physician visits.) The follow-up physician visit cost figure in Table 2 was also applied to the 17 follow-up visits in the Urgent Care Center. Costs for the 20 ENT visits were not included in our calculations. Since 13 of these visits occurred in episodes in which the physician was the initial provider, this omission tends to bias our cost comparisons in favor of physicians.

TABLE 2. Provider-Time–Related Costs per Visit

	MD	NP	NP with MD Consultation
Otitis media			
Initial	$7.91	$3.70	$7.61
Follow-up	$6.38	$2.32	$6.23
Sore throat			
Inital	$6.55	$3.61	$7.52
Follow-up	$9.35	$2.62	$6.53

Ancillary Service and Drug Costs

The cost of drugs was figured as the acquisition costs to the plan, plus a $2 service charge. This is the formula used by the plan in its own billing system. Similarly, costs for x-ray services were equated with the rates billed for these services to fee-for-service customers. Laboratory costs were equal to charges paid by the plan to outside laboratory facilities.

Episode Cost Calculations

The costs for each episode were computed by adding the relevant cost figure in Table 2 for each visit in the episode to the cost of ancillary services and drugs. For example, the cost of an otitis media episode consisting of an initial nurse practitioner visit, two follow-up physician visits, and $5.00 in drug and ancillary costs would be $3.70 + 2 ($6.38) + $5.00 = $21.46. Mean cost figures for episodes grouped by problem type and provider type at the initial visit were then calculated.‡

Analysis of EMCRO Data

As part of the Columbia EMCRO project,[21] questionnaire data on patient-reported problem status were gathered for all visits to the Department of Pediatrics

‡ Standard deviations of the mean cost figures were also calculated and used in the statistical tests reported. However, these standard deviations do not precisely measure (and probably understate) the variability in episode costs for several reasons. They only reflect variability in numbers of visits and in ancillary and drug costs. They do not include variability in provider time inputs at each visit, covariances between provider time inputs at different visits in an episode, or covariances between time inputs and drug or ancillary costs. These deficiencies are to some extent unavoidable, since our methodology did not entail observation of time inputs at each visit of every episode studied. (The cost of obtaining such observational data would have been considerable.) Consequently, statistical tests relating to our mean episode cost figures must be viewed as suggestive rather than definitive.

during a two-week period in mid-1974. Problem status was described in terms of amount and intensity of pain and discomfort in the previous 24 hours, degree of activity limitation in the previous three days and level of anxiety about the problem. One week telephone follow-ups were also made for a 50 per cent random sample of patients to determine change in problem status subsequent to the visit. Tabulations of these data for persons with initial visits for otitis media and sore throat are presented below. The purpose of these tabulations was to determine whether patient-reported problem severity and changes in problem status differed between physician and nurse practitioner patients.

Findings

Selected patient and problem characteristics for the 409 otitis media and 390 sore throat episodes, classified by type of provider (physician or nurse practitioner) at the initial visit, are described in Table 3. For nurse practitioner episodes, data are also shown separately for cases with and without physician consultation at the initial visit. In both problem categories, about 10 per cent of initial nurse practitioner visits involved physician consultation. The characteristics of the patients and problems seen by physicians and nurse practitioners appear to be fairly similar. Physician patients with otitis media are slightly younger, but this is not the case for sore throat. The percentages of episodes with multiple problems are about equal, although, as might be expected, nurse practitioner episodes with physician consultation show higher percentages of multiple problems. The relative frequency of serous otitis media is insignificantly higher in nurse practitioner episodes.

Various measures of service utilization for episodes are shown in Table 4. These figures indicate that nurse practitioner patients with otitis media had about twice as many laboratory tests as physician patients, while physician sore throat patients had almost twice as many follow-up visits as nurse practitioner patients. All other differences between physician and nurse practitioner patients are small in relative terms. (This conclusion also holds for x-ray use, which is not shown in Table 4 because it occurred in only seven of the 799 study episodes.)

TABLE 3. Patient and Problem Characteristics for Episodes

	Otitis Media				Sore Throat*			
	MD	NP	NP Only	NP With MD	MD	NP	NP Only	NP With MD
Number of episodes	176	233	204	29	185	205	184	21
Mean patient age (years)	4.85	5.55*	5.56	5.48	7.53	7.52	7.52	7.57
Per cent of patients under 3 years	33.5	19.7*	20.6	13.8	11.4	10.7	10.3	14.3
Per cent with secondary Dx	31.3	39.9	37.3	58.6	25.4	22.4	20.1	42.9
Per cent with serous otitis media	15.3	21.5	21.6	20.7	—	—	—	—

* MD–NP difference is significant at the 0.05 level (two-tailed).

TABLE 4. Utilization Statistics for Episodes

	Otitis Media				Sore Throat			
	MD	NP	NP Only	NP With MD	MD	NP	NP Only	NP With MD
Per cent with follow-up visits	28.4	33.9	33.3	37.9	12.4	6.8*	6.0	14.3
Average number of follow-up visits	0.59	0.57	0.56	0.66	0.15	0.08	0.08	0.14
Average number of MD follow-ups†	0.41	0.26	0.25	0.35	0.12	0.05	0.06	0.05
Average number of NP follow-ups†	0.15	0.25	0.25	0.26	0.03	0.03	0.02	0.10
Average number of NP/MD follow-ups†	0.03	0.06	0.06	0.03	0.00	0.00	0.00	0.00
Per cent with lab tests	17.6	32.2	31.9	34.5	98.4	98.5	98.4	95.2
Average number of lab tests	0.26	0.45*	0.45	0.41	1.42	1.29	1.26	1.52
Per cent with prescriptions	84.7	91.8*	90.7	100.0	59.5	58.5	58.2	61.9

* MD–NP difference is significant at the 0.05 level (two-tailed).
† Significance of MD–NP difference was not computed for this item.

Conversely, differences in mean episode costs, which are reported in Table 5, are fairly large. For otitis media episodes, costs for patients of nurse practitioners are 18 per cent below costs for physician patients; for sore throat episodes, the differential is 25 per cent. Since differences in service utilization are minor, percentage differences in mean ancillary costs are more modest. Nurse practitioner sore throat patients' ancillary costs are 9 per cent lower than those for physician patients, while nurse practitioner otitis media patients' ancillary costs are 14 per cent higher. Thus, the major source of the overall savings for nurse practitioner patients is in the provider-time-related costs. For both problem categories, nurse practitioner provider-time-related costs are about 40 per cent lower.

It has already been noted that the visit time estimates used in the cost calculations were based on small numbers of observed visits, but it is doubtful that sampling errors in these estimates greatly affected the qualitative nature of our findings. While the

TABLE 5. Episode Costs

	Otitis Media				Sore Throat			
	MD	NP	NP Only	NP With MD	MD	NP	NP Only	NP With MD
Mean provider time Related costs ($)	11.05	6.79*	6.24	10.66	7.75	4.59*	4.18	8.51
Mean ancillary costs ($)	7.17	8.19	7.95	9.87	7.89	7.21	7.15	7.71
Mean total costs ($)	18.22	14.98*	14.19	20.53	15.64	11.80*	11.33	15.92

* MD–NP difference is significant at the 0.05 level (two-tailed).

follow-up visit time estimates are especially tentative, the frequency of follow-up visits in the study episodes is low. Moreover, because the cost differences in Table 5 are large, errors in the time estimates would have to be relatively large to explain these differences.

For example, based on the data in Tables 1, 3 and 4, the average physician contact time for a physician otitis media episode was 12.513 minutes (9.3 minutes + 0.41 × 7.5 minutes + 0.03 × 4.6 minutes), while physician contact time for a nurse practitioner otitis media episode was 2.799 minutes (0.26 × 7.5 minutes + [29/233] × 4.6 minutes + 0.06 × 4.6 minutes). To eliminate the mean episode cost differential of $3.24 ($18.22 − $14.98), one would have to reduce the difference in these times by 3.81 minutes, that is, from 9.714 minutes (12.513 minutes − 2.799 minutes) to less than 5.904 minutes.§ This implies that our estimates of physician contact time inputs would have to be at least 64.5 per cent ($\frac{3.81}{5.904} \times 100$) above the true time input figures to explain fully the $3.24 cost differential reported in Table 5. Similar calculations for sore throat episodes show that our estimates of physician contact time inputs would have to be at least 182.2 per cent above the true figures to explain fully the reported cost differential of $3.84. Using the same approach, it can be shown that fully explaining the mean episode cost differentials by underestimates of nurse practitioner contact times would require that the true nurse practitioner time inputs be roughly twice as large as our estimates.

Before concluding that the results presented in Table 5 represent efficiency differences, two other possible explanations should be noted. First, since patients were not randomly assigned to physicians or nurse practitioners, it is possible that physician patients tended to have severe problems. Second, the cost savings achieved by nurse practitioners could simply reflect the provision of lower quality or less effective care.

Tabulation of data from the Columbia EMCRO project (Table 6) does not support either of these explanations. Patient-reported problem severity (i.e., discomfort and activity limitation) at the initial visit did not differ significantly by provider type. Similarly, no significant difference was found in the change in problem status one week after the initial visit. Moreover, while the number of visits included in these tabulations is small, the results in Table 6 are supported by other evidence. In particular, the hypothesis of differences in problem severity by provider type has been rejected in an earlier analysis[22] of Columbia EMCRO data for all types of problems combined. Our own data on patient and problem characteristics (Table 3) also support this conclusion. The hypothesis of provider-type difference in quality or effectiveness were tested in the earlier EMCRO analysis,[22] with data on change in problem status (over one week and one month) and patient satisfaction for all problem types combined, and was rejected.

Summary and Discussion

Concerns with ambulatory health care costs, and with the potential for reducing these costs and improving productivity through the use of NHPs, have been widely discussed. Most previous assessments of this potential have used the visit as the unit of comparison. Important limitations of this method include its failure to account for variations in the use of ancillary and related services, and variations in practice patterns regarding referral and follow-up visits. (Note that in the current study significant physician–nurse practitioner differences were found in the use

§ The 3.81 figure is obtained by dividing $3.24 by $0.85, which is the multiplication factor for T_{MD} in our cost formula.

TABLE 6. Initial Problem Status and Change in Problem Status in Otitis Media and Sore Throat Episodes

	Otitis Media				Sore Throat*	
	MD	NP†	NP Only	NP‡ With MD	MD	NP Only
Per cent of episodes with:						
No discomfort	16.4	11.8	9.1	28.6	15.4	—
Some discomfort	29.1	21.6	22.7	14.3	19.2	31.4
Moderate discomfort	38.2	45.1	43.2	57.1	46.2	51.4
Severe discomfort	16.4	21.6	25.0	—	19.2	17.1
(N)	(55)	(51)	(44)	(7)	(26)	(35)
Per cent of episodes with:						
No activity limitation	37.0	42.0	41.9	42.9	30.8	19.4
Some activity limitation	31.5	26.0	25.6	28.6	34.6	36.1
Moderate activity limitation	27.8	28.0	27.9	28.6	30.8	44.4
Severe activity limitation	3.7	4.0	4.7	—	3.8	—
(N)	(54)	(50)	(43)	(7)	(26)	(36)
Per cent of episodes with problem status:						
Resolved	27.8	53.8	53.8	—	42.9	75.0
Better	61.1	46.2	46.2	—	35.7	25.0
Unchanged	11.1	—	—	—	21.4	—
(N)	(18)	(13)	(13)	—	(14)	(38)

* No NP with MD initial visits for sore throat were reported.
† None of the MD–NP differences shown here are significant at the 0.05 level.
‡ No follow-up data or change in problem status were reported for these visits.

of laboratory tests and drugs for otitis media patients and in follow-up visit rates for sore throat patients.) The comparison of outcome or effectiveness measures for single visits is also problematic. A methodology using the person-year of care overcomes these limitations, but because it involves examination of all services received over a defined period, it is subject to other confounding influences (e.g., random variation in the incidence of health problems, productivity and quality variations among *all* provider organizations serving the study population).

The conceptual advantages of analyses based on episodes of care have been noted by several researchers. In this study, a practical methodology for episode-based comparisons was developed and employed to assess the efficiency of physician and nurse practitioner care for two common pediatric conditions, otitis media and sore throat. (According to our observational study, these two conditions accounted for more than 25 per cent of all pediatric visits.) The results indicated that nurse practitioner care was less costly and that this cost differential could not be attributed to physician–nurse practitioner differences in patient characteristics, problem severity, or effectiveness or quality of care. The reliability of this result is, of course, dependent upon the limitations of our data and methods noted previously. The potential for generalizing this result is also limited to settings with similar organizational and cost characteristics. The methodology, however, can be generally applied. It makes use of existing financial and ambulatory care management information systems, supplemented by a small-scale observational study. The most complex element is the development and application of decision rules to link individual visits and ancillary services systematically into episodes of care. However, previous research[23] has demonstrated that such criteria can be developed empirically and/or by a consensus process among health care providers.

The utility of episode-based cost comparisons (such as that reported in this article) can be seen at several levels. At the clinical level, triage decisions regarding the type of provider to be seen for a problem can be influenced by data such as those presented here. In addition to examining provider differences, this methodology, linked with quality measurements, can be applied to comparisons of different clinical approaches to treatment and follow-up, as well as to the influence of patient education efforts. At the organizational level, productivity information is influential in staffing and salary decisions, and in decisions regarding facilities and the organization of clinical units. On the broader social level, these data help to focus attention on questions of efficiency and effectiveness when care is provided under different modes of organization and financing, and to different population groups.

References

1. Record JC, et al. Cost effectiveness of physician's associates: Kaiser-Permanente experience. American Economic Association Annual Meeting, Dallas, Texas, 1975.

2. Smith KR, et al. An analysis of the optimal use of inputs in the production of medical services. J Hum Resour 1972;7:208.

3. Hoekelman RA. What constitutes adequate well-baby care? Pediatrics 1975;53:313.

4. Schiff DW, et al. The pediatric nurse practitioner in the offices of pediatricians in private practice. Pediatrics 1969;44:62.

5. Charney E, Kitzman H. The child-health nurse (pediatric nurse practitioner) in private practice: A controlled trial. N Engl J Med 1971;285:1353.

6. Denari GA, Katz HP. Periodic quality assessment of a revised well-baby care program in a prepaid multispecialty group practice. Ambulatory Pediatrics Association 14th Annual Meeting, Washington, D.C., 1974.

7. Kane RL, et al. Effects of adding a medex on practice costs and productivity. J Community Health 1978;3:216.

8. Romm J, et al. Survey and evaluation of the physician extender reimbursement study: Cost effectiveness. American Public Health Association Annual Meeting, Los Angeles, 1978.

9. Morris SB, Smith DB. An evaluation of the potential impact of physician extenders on primary care practices. Second International Congress of World Federation of Public Health Association, Halifax, Nova Scotia, 1978.

10. Steinwachs DM, et al. The role of new health practitioners in a prepaid group practice: Changes in the distribution of ambulatory care between physician and nonphysician providers of care. Med Care 1976;14:95.

11. Chambers LW, et al. A controlled trial of the impact of the family practice nurse on volume, quality, and cost of rural health services. Med Care 1977;15:971.

12. Spector R, et al. Medical care by nurses in an internal medicine clinic: Analysis of quality and its costs. JAMA 1975;232:1234.

13. Lewis CE, Resnick B. Nurse clinics and progressive ambulatory patient care. N Engl J Med 1967;277;1236.

14. Scitovsky AA. An index of the cost of medical care—a proposed new approach. The economics of health and medical care. Proceedings of the Conference on the Economics of Health and Medical Care. Ann Arbor: University of Michigan, 1964.

15. Record JC. An episode approach to utilization, costs and effectiveness in health care. Unpublished paper, Kaiser Foundation Health Services Research Center, Portland, 1978.

16. Moscovice I. A framework for analyzing resource use in ambulatory care settings. Med Care 1977;15:1024.

17. Tompkins RK, et al. The effectiveness and cost of acute respiratory illness medical care provided by physicians and algorithm-assisted physician's assistants. Med Care 1977;15:991.

18. U.S. Congress, Congressional Budget Office. Physician Extenders: Their Current and Future Role in Medical Care Delivery. Washington, D.C.: U.S. Government Printing Office, 1979.

19. West SK, et al. Drug utilization review in a health maintenance organization. The Johns Hopkins Medical Institutions. Final report submitted to the National Center for Health Services Research, DHEW, 1977.

20. Skinner EA, et al. Productivity analyses of ambulatory care providers: A method based on observational and information system data. Proceedings of the Annual Conference of the Society for Advanced Medical Systems, Washington, D.C., 1980.

21. Mushlin AI, Appel FA. Testing an outcome-based quality assurance strategy in primary care. Med Care 1980;18(suppl).

22. Levine DM, et al. The role of new health practitioners in a prepaid group practice: Provider differences in process and outcomes of medical care. Med Care 1976;14:326.

23. Steinwachs DM, Yaffe R. An approach to the review of utilization and quality of ambulatory care. Joint Meeting of the Operations Research Society of America and The Institute of Management Science, San Juan, Puerto Rico, 1974.

MEDICAL CARE
February 1982, Vol. XX, No. 2

Costs and Outcomes of Medical Intensive Care

MARK R. CHASSIN, M.D., M.P.P., M.P.H.

Survival, functional status and total six-month medical care costs were measured for 489 consecutive medical intensive care patients. In addition, the study sought to determine if it was possible from data available on admission to identify subgroups of patients with extremely low survival rates. Of the study patients, 86 per cent survived the initial admission, and 72 per cent survived six months. Daily intensive care unit (ICU) charges averaged $1,120 per patient, while total six-month charges per patient averaged $14,577. Age, initial systolic blood pressure, preadmission functional status and diagnosis were all significantly and independently associated with survival. The subgroup with the lowest six-month survival rate (14 per cent) comprised cancer patients with poor preadmission functional status. This study did not identify a subgroup of patients with a survival rate low enough to justify systematic exclusion from intensive care. Controlling ICU costs by excluding patients with very low expected survival may not be possible.

INTENSIVE CARE has spread widely in the last 20 years, yet little is known about its overall effectiveness. The percentage of nonfederal, short-stay hospitals reporting any facilities for intensive care has increased from less than 10 per cent in 1958 to more than 74 per cent in 1977 (Fig. 1).[1,2] While intensive care apparently contributes dramatically to saving individual lives that would previously have been lost, it also generates levels of cost not previously attained. Moreover, this technology contributes to the increasing cost of hospital care. One conservative estimate suggests that intensive care accounts for almost 20 per cent of total hospital costs.[3]

Considering the level of investment in intensive care and the speed with which it has been rising, it is important to define carefully which patients most benefit from it, the degree to which they benefit and the cost of these additional benefits. In 1973, Griner called attention to the lack of research evaluating intensive care.[4] Six years later, Ebert justifiably asserted that intensive care units (ICUs) have proliferated despite a lack of clearly defined indications for patient admissions.[5] This paucity of evaluative research is even more conspicuous in light of mounting pressures to contain the growth of hospital costs.

Of all the varieties of intensive care, medical intensive care has been the least studied. Much work has been done on coronary intensive care,[6–11] and there are

From the Department of Medicine, Center for the Health Sciences, University of California at Los Angeles, and The Robert Wood Johnson Foundation Clinical Scholars Program, University of California at Los Angeles.

Supported by a grant from the Robert Wood Johnson Foundation, Princeton, New Jersey. The opinions, conclusions, and proposals in the text are those of the author and do not necessarily represent the views of The Robert Wood Johnson Foundation.

Requests for reprints: Mark R. Chassin, M.D., The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406.

0025-7079/82/0200/0165/$01.25 © J. B. Lippincott Co.

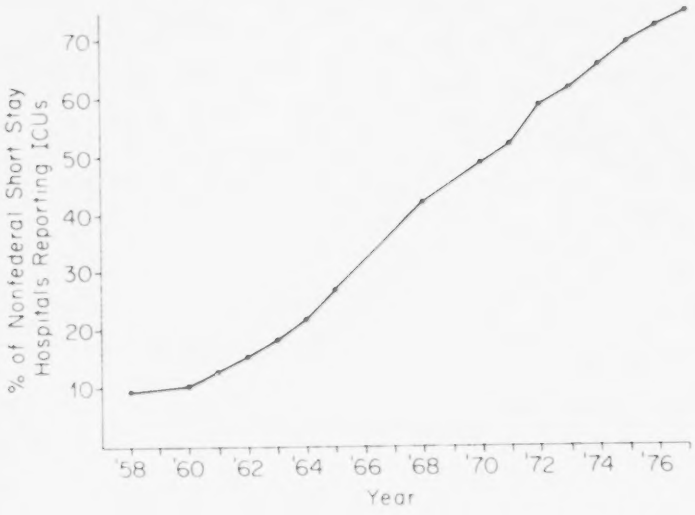


FIG. 1. Rise in percentage of nonfederal, short-stay hospitals reporting availability of intensive care units.
Sources: 1971–77 data[1]
1958–70 data[2]

some published evaluations of surgical intensive care.[12–16] More recently, neonatal intensive care has undergone extensive scrutiny; a randomized clinical trial has been conducted,[17] and the entire experience has been reviewed.[18]

In contrast, literature on the effectiveness and costs of medical intensive care is scarce. Several early studies provided the diagnostic categorization of medical intensive care patients.[19–21] Spagnolo and his coworkers reported a comprehensive set of in-hospital mortality statistics for a general medical intensive care unit (MICU).[22] But only three studies have been reported in which comparisons were made between patients treated in an MICU and comparable patients treated in general ward settings; the first study showed no effect on the survival of patients with pulmonary edema after the opening of an MICU,[23] and the second showed decreased mortality among patients receiving mechanically assisted ventilation following the opening of a respiratory intensive care unit (RICU).[24] The validity of the latter data is, however, open to question because it is not clear that the indications for mechanical ventilatory assistance remained the same in both periods under study. The third study showed no effect on mortality among patients treated for drug overdose after the opening of an RICU, although the rate of complications decreased.[25] The remainder of the published reports give either in-hospital mortality data only or follow-up survival data on a limited range of patients.[26–35] Only one study has reported even limited follow-up data on a general medical intensive care population.[36] Even less information has been published on the costs of medical intensive care,[23,25,35–37] though recently some attention has been given to high-cost patients.[38,39]

In considering how to moderate ICU costs, three different approaches are conceptually apparent. One might try to limit the number of admissions to ICUs by using criteria that more strictly define the appropriate use of intensive care. Second, one might attempt to reduce the length of stay by discharging patients sooner. Finally, one might try to reduce the average cost of a day in an ICU. The present study addresses the feasibility of one variation on the first theme.

One approach to ICU cost control that has received some attention in the litera-

ture is to develop criteria for ICU admission and discharge that would exclude from ICUs those patients classified as "hopelessly ill" and for whom "selective limitation of therapeutic measures" is planned[40] or those patients whose short-term prognosis is poor and for whom only "ordinary therapeutic efforts" are planned.[35] Criteria developed so far have been entirely subjective and rely on the attending physician's clinical judgment. There has been little attempt to be explicit about what is meant by "poor short-term prognosis" or about what types of patients might fall into this category.

The study of 489 consecutive medical intensive care patients described here was designed to define one hospital's medical intensive care patients demographically and clinically, and to measure their survival, their functional status and the costs of their care. In addition, the study sought to test the hypothesis that at the time of ICU admission, data are available from the medical record that would allow one to identify subgroups of patients with an especially low likelihood of survival.

Methods

All admissions to the RICU and MICU at a major university teaching hospital in Los Angeles were reviewed for the study period, which extended from January 1, 1977 through July 31, 1977. During this period, 566 patients accounted for 671 admissions to these two ICUs. Ninety-eight per cent of the inpatient medical records were obtained. Of these, 11.6 per cent (64) were rejected from the study for one of four reasons: the patient was a surgical patient, usually admitted for treatment of a postoperative complication (8.0 per cent); the patient was a Coronary Care Unit patient, being attended by CCU staff (2.5 per cent); the patient was transferred to another hospital before leaving the ICU (0.7 per cent); or the patient left the ICU against medical advice (0.4 per cent). The remaining 489 patients form the sample that was studied further.

The RICU at the study hospital comprises eight beds and is staffed by a team of physicians including house staff, a pulmonary fellow and one attending physician from the pulmonary division. The team is separate from the general medicine ward teams. All admissions to the RICU were approved by the pulmonary fellow during the time of the study. The MICU also consists of eight beds. Patients in this unit were managed by the general medicine ward teams during the study period. Admissions to this unit were usually arranged by the senior admitting resident in the emergency department. Both ICUs had similar nurse staffing patterns, including the capacity for one-to-one nursing when indicated, and both had access to similar equipment. In the event that both the RICU and the MICU were full, a medical intensive care patient might be admitted to a nonmedical ICU. Some of these patients would be transferred to either the RICU or the MICU, if medically appropriate, when a bed became available. Such patients are included in the present study. Patients receiving all of their medical intensive care in nonmedical ICUs are, however, not included. The term "medical ICUs" will be used in the remainder of this report to include both the RICU and the MICU of the study hospital.

Each hospital record was reviewed for demographic and diagnostic data, as well as for clinical and laboratory data relevant to severity of illness. In patients with multiple diagnoses, it was uniformly possible to select the one major disease process that was most life-threatening and for which the ICU admission was required. In addition, medical and surgical procedures and physician visits were tabulated by type of service and according to whether the patient was in the ICU or on a ward when the service was rendered.

Each patient's hospital bill was obtained and subdivided by kind of charge and by the location of the patient at the time the charge was incurred. A list of medical procedure fees was obtained from the hospital division that prepares the bills for such procedures. The following pricing decisions for charges for physician inpatient services were made to reflect an average or low-average level of charge for the indicated service within the geographic area served by the study hospital. Each admission history and physical examination by an attending physician, and each new consultation was valued at $120; each follow-up visit by an attending physician or consultant, as documented by a signed note in the medical record, was valued at $30. Surgical fees were determined by reference to the 1964 California Relative Value Studies. Each unit was assigned a weight of $55.

Next, each patient or surviving family member was interviewed by telephone. The interviews were conducted from six to 20 months following discharge (median 12 months). Data were obtained at that time concerning patients' functional status six months after discharge and their functional status during the time preceding the initial hospitalization. The questions used to assess functional status were adapted from the physical limitations scale used in the Rand Health Insurance Study.[41] In addition, patients were asked about their utilization of medical services during the six months after discharge; data were collected on the frequency of physician visits, laboratory and x-ray examinations, other tests and procedures, drugs and equipment, treatments by nonphysician medical personnel, rehospitalization and residence in an extended care facility.

Finally, when patients or family members were unable to provide cost data for these follow-up services, as was most often the case, each local physician, hospital and nursing home was contacted and the charges for their services recorded. A survey was conducted of three pharmacies in the area to price the drug regimens followed by study patients. The prices charged by the study hospital's pharmacy, a private hospital pharmacy, and a popular discount pharmacy were averaged to arrive at a single price for each drug used by these patients. The regimen described by the patient in the interview was then priced accordingly.

No distinction is drawn in this report between charge and cost because of the difficulty in obtaining data on the true costs of medical services. In most cases charge data are adequate estimates of true costs. It is important to realize, however, that in certain areas, particularly inpatient ancillary services, charges may exceed true costs.

Of the 489 initial hospital bills, all but six were obtained for analysis. During the six months following discharge, the survivors had a total of 344 readmissions to hospital. Only four of the 207 bills for readmission to the study hospital were not obtained, and only nine of the 137 bills from other hospitals were not available. No difficulty was encountered in obtaining charge data from area physicians or nursing homes. Survival data were obtained on 97 per cent of the sample. Functional status data were not obtained for 39 patients (8 per cent). Ten patients (2 per cent) refused to be interviewed. In 29 cases (6 per cent) neither the patient nor any close family member or friend was available for the interview, but survival data were obtained for 12 of these cases.

Results

A histogram of the group's age distribution is presented in Figure 2; the mean age was 54 years. Fifty-four per cent of the patients were female. Table 1 presents the frequency of the major diagnostic categories responsible for the medical ICU admissions of these patients and the most common subgroups within the first five

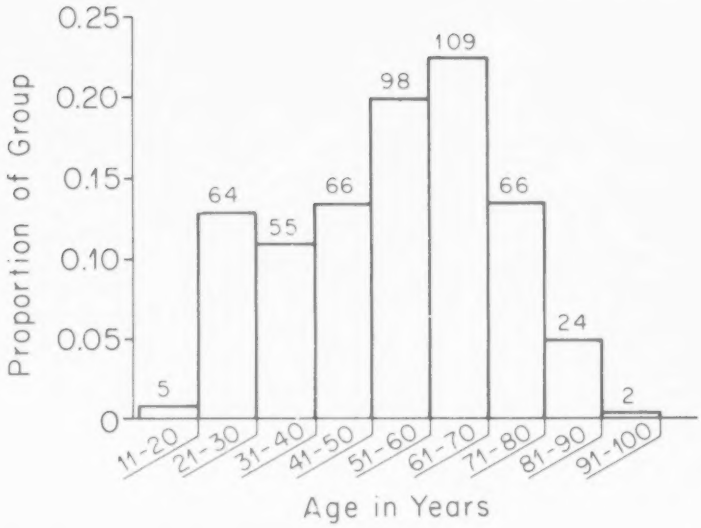


FIG. 2. Histogram of age distribution of medical intensive care population.

Numbers above histogram bars refer to numbers of patients in each age group.

categories. Each category represents a broad range of diseases, and only three homogeneous diagnostic groups each contain 5 per cent or more of the sample: chronic obstructive pulmonary disease (COPD), congestive heart failure (CHF) and overdose. These medical ICUs, therefore, treat a wide spectrum of illnesses, each of which represents a very small proportion of the total.

The average length of ICU stay was 5.1 days, with a range of 1 to 95 days. Patients who stayed only one day in the ICU composed 27 per cent of the sample, patients who stayed two days made up an additional 22 per cent, while another 10 per cent stayed more than ten days. The total hospital stay was similarly widely distributed, with a mean of 13.5 days and a range of 1 to 120 days.

Of the 489 study patients, 70 died during the initial hospital admission, accounting for an in-hospital survival rate of 86 per cent. At six months following discharge, 134 patients had died and 17 had been lost to follow-up, accounting for a survival rate of 72 per cent at six months. Figure 3, showing the survival curve for the group as a whole using the life table method of analysis, reveals a one-year survival rate of 66 per cent.

The functional status questionnaire permitted division of the group into six

TABLE 1. Major Diagnostic Categories and Selected Subgroups

Category	Number of Patients	Per Cent
Cardiovascular	110	22.5
Congestive heart failure	38	7.8
Hypertension	16	3.3
Pulmonary	106	21.7
COPD	48	9.8
Pneumonia	14	2.9
Gastrointestinal	73	14.9
Duodenal ulcer	15	3.1
Gastritis	13	2.7
Inflammatory bowel disease	10	2.0
Cancer	59	12.1
Solid tumors	37	7.6
Hematologic	22	4.5
Neurologic	36	7.4
CVA (cortical)	14	2.9
Overdose	28	5.7
Renal	20	4.1
Rheumatologic	19	3.9
Diabetes mellitus	15	3.1
Other	23	4.7
Total	489	100.1

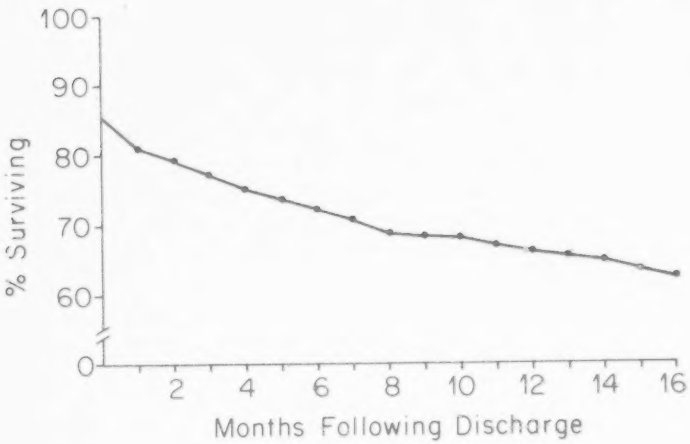


FIG. 3. Survival curve of medical intensive care patient population.
This curve was constructed using the life table method.

classes based on the ability to perform certain activities. Table 2 presents the distribution of functional status classes before admission and at six months after discharge. Considering only survivors, the two distributions are virtually identical. Of the 450 patients for whom functional status data were available, 42 per cent reported that they were in the same functional class six months after discharge as they had been before hospitalization. Only 8 per cent reported an improved functional status, while 20 per cent reported worsened functional capacities. The remaining 30 per cent were dead.

TABLE 2. Distribution of Functional Status

Class*	Before Hospitalization (%)	Six Months After Discharge	
		Whole Group (%)	Survivors (%)
5	29	18	26
4	13	11	16
3	21	13	18
2	21	15	22
1	6	5	7
0	10	8	11
dead	—	30	—

* Definitions of functional classes:

Class 5: Able to perform vigorous activities such as running or lifting heavy objects.

Class 4: Unable to perform vigorous activities, but able to walk several blocks.

Class 3: Unable to walk several blocks, but able to walk one block.

Class 2: Able to walk alone, but not as far as one block.

Class 1: Unable to walk alone, but able to eat, dress, and bathe without assistance.

Class 0: Unable to walk or perform self-care activities without assistance.

Survival

Table 3 presents the univariate relationships between survival and each of many patient characteristics for which statistically significant associations were demonstrated. The chi-square test was used to evaluate the statistical significance of these associations. Diagnosis was the most effective characteristic for distinguishing those subgroups of patients with very high survival rates from those with very low ones. Patients with cancer, including both solid tumors and hematologic malignancies, who were sick enough to be treated in these medical ICUs fared the worst. No other way of categorizing the patient population produced a group with a lower six-month survival rate than the 24 per cent experienced by the cancer patients. Nor did any other method of classification produce a group with a higher rate than the patients with hypertension, 93 per cent of whom survived six months.

TABLE 3. Univariate Associations with Survival

Characteristic	Number of Patients	Six-month Survival (%)
TOTAL GROUP	472	71.6
Sex		
Male	213	66.7
Female	259	75.7*
Age		
30 and under	65	89.2
31–50	112	76.8
51–70	204	66.2
over 70	91	64.8†
Initial systolic blood pressure		
100 mmHg and less	65	58.5
over 100 mmHg	407	73.7‡
Preadmission functional status		
Class 0	45	44.4
Class 1	27	59.3
Class 2	96	63.5
Class 3	94	70.2
Class 4	59	81.4
Class 5	129	81.4†
Preadmission employment status		
Full-time	105	83.8
Part-time	29	58.6
Housewife	43	81.4
Student	20	85.0
Not working	253	62.8†
Initial mental status (excluding overdoses)		
Alert	343	77.3
Lethargic	50	58.0
Stuporous	31	38.7
Semi-comatose and comatose	20	30.0†
Diagnosis		
Hypertension	14	92.9
Overdose	27	92.6
Diabetic ketoacidosis	10	90.0
COPD	47	85.1
Upper gastrointestinal bleeding	60	80.0
Pneumonia	14	78.6
Stroke (cortical)	14	71.4
Chronic renal failure	16	62.5
Congestive heart failure	37	59.5
Cancer	59	23.7†

* = $p < 0.05$.
† = $p < 0.005$.
‡ = $p < 0.025$.

Multivariate analysis confirmed and elucidated further the relationships described in Table 3. The patient population was divided into eight subgroups on the basis of age, initial systolic blood pressure and preadmission functional status; each variable was dichotomized. These particular variables were chosen for further analysis because of their strong univariate associations with survival. For each vari-

able, the dividing line was chosen so as to maximize the difference between the two subgroups so created. Comparison of the six-month survival of these subgroups revealed a uniform trend. Patients who were older, who were admitted to the ICU in shock or who had low preadmission functional status did less well than their respective counterparts. This was established by comparing the effect of each variable on survival while controlling for the other two. The subgroup with the best six-month survival—those under 50 years of age, admitted with systolic blood pressures above 100 mmHg, and with preadmission functional status of class 4 or 5—experienced a survival rate of 90 per cent. In the subgroup with the lowest survival—those over 50, with admitting systolic blood pressures of 100 mmHg or less, and in the lowest four functional classes prior to admission—only 39 per cent were alive six months after discharge. Even this subgroup of 18 patients, however, had a better survival experience than did the patients with cancer.

The independent association of these three variables with survival is noted also when one looks within diagnostic groups. Thus, for patients with congestive heart failure, those who were 60 years of age or younger experienced a six-month survival rate of 85 per cent, while those older than 60 experienced a rate of only 50 per cent. Similarly, those CHF patients who reported preadmission functional status in classes 3–5 survived at a rate of 71 per cent, while those in the lower functional classes survived at a rate of 47 per cent. Because of small sample sizes, neither of these large differences is statistically significant.

For patients admitted to these medical ICUs with upper gastrointestinal bleeding, 84 per cent of those 60 years old or younger survived six months, while 73 per cent of those over 60 survived. Of those admitted with systolic blood pressures initially over 100 mmHg, 82 per cent lived through the follow-up period, but only 71 per cent of those admitted with initial systolic pressures of 100 mmHg or less were alive six months after discharge. Further, 88 per cent of those in the highest two functional classes survived, while only 69 per cent of those in functional classes 0–3 lived. None of these differences is statistically significant.

Finally, for patients with cancer, those in functional classes 4–5 had a six-month survival rate of 43 per cent, but those in classes 0–3 experienced a rate of only 14 per cent; this difference is statistically significant at the 5 per cent level. The subgroup of 36 cancer patients with poor preadmission functional status experienced the lowest survival rate of any subgroup identified in the study.

Costs

The average daily ICU charge for this patient population exceeds $1,000 per day, as seen in Table 4. It can readily be appreciated from these data that the largest single component of the ICU segment of the hospital bill is represented by hospital services, which constitute 44 per cent of the bill; laboratory charges account for 23 per cent, and professional fees make up 14 per cent. If one adds the charges for hospital care in the follow-up period ($3,900) to those for the initial hospital stay ($9,601), it is clear that hospital care accounts for 93 per cent of average total six-month charges.

There was little variation in these average charges by age, initial systolic blood pressure, preadmission functional status or initial mental status. There was significant variation by diagnosis; diabetic ketoacidosis had the lowest six-month average ($9,899) and chronic renal failure the highest, with a value of $31,668. Survivors averaged $12,860, while nonsurvivors averaged $19,838 for the six months. Both of these differences are statistically significant ($p < 0.005$, Student's t-test). Interestingly, the average charges accumulated by

those nonsurvivors who died during the initial hospitalization ($19,812) were virtually identical to those incurred by the nonsurvivors who died during the follow-up period ($19,868).

Figure 4 shows the extent to which resources for intensive care patients are spread unevenly across the patient population; 17 per cent of the patients consumed half of the total resources. A subgroup of 32 patients (7.4 per cent), each of whom incurred more than $40,000 in total six-month charges, was designated as the high-cost subgroup and examined more closely. This subgroup accounted for 31 per cent of the total resources. The average charge in this subgroup was $62,736, and the range was from $40,432 to $135,547.

Table 5 compares selected characteristics of the high-cost subgroup with those of the group as a whole. Mean age, functional status distribution and proportion of severely hypotensive patients are similar between the two groups. There are slightly more males in the high-cost subgroup and more patients admitted with an abnormal mental status. As expected, the ICU length of stay for these patients was markedly longer than that for the entire group. There were relatively fewer patients with cancer in the high-cost subgroup, but relatively more with chronic renal failure and rheumatologic conditions. Survival was modestly, but not dramatically, reduced.

Thus while there are some differences between the high-cost subgroup and the medical intensive care population, in general the high-cost subgroup is broadly representative of the whole. It is not narrowly confined to one age group, functional class or diagnostic category, nor does it have a uniformly poor outcome. The only striking difference, length of stay, was clearly dictated by the selection of the patients with the highest costs. The close correlation between length of stay and cost is clear; the longer one stays in a hospital, on average, the higher the cost will be. This association, however, is of limited usefulness in identifying patients who are likely to use large amounts of resources before those resources are consumed. It was not possible in this study to identify any characteristics available on admission that would identify those patients who would later enter the high-cost subgroup.

TABLE 4. Analysis of Average Total Six-Month Charges

Type of Service	Total ($)	Cost Per Day ($)
Initial Hospitalization		
ICU hospital services*	2,485	489
ICU inhalation therapy	317	62
ICU pharmacy†	452	89
ICU X-ray‡	317	62
ICU laboratory§	1,318	259
ICU professional fees‖	809	159
Average total ICU charges	5,698	1,120
Average total ward charges	3,903	
Average total initial hospitalization	9,601	
Average charges for rehospitalization during six-month follow-up period	3,900	
All other charges during six-month follow-up period	1,076	
Average total follow-up charges¶	4,976	
Average total six-month charges per patient	14,577	

* Includes charges for room and board, hemodialysis, emergency and operating rooms, supplies, and physical therapy.

† Includes charges for all medications and intravenous fluids.

‡ Includes charges for diagnostic radiology, ultrasound (except echocardiography), nuclear medicine and radiation therapy.

§ Includes all clinical laboratory charges as well as charges for electrocardiography, echocardiography, blood gas analyses, outside laboratory tests, and pathology.

‖ Includes charges for visits, consultations and procedures, including those for anesthesia.

¶ Includes nonsurvivors in a denominator.

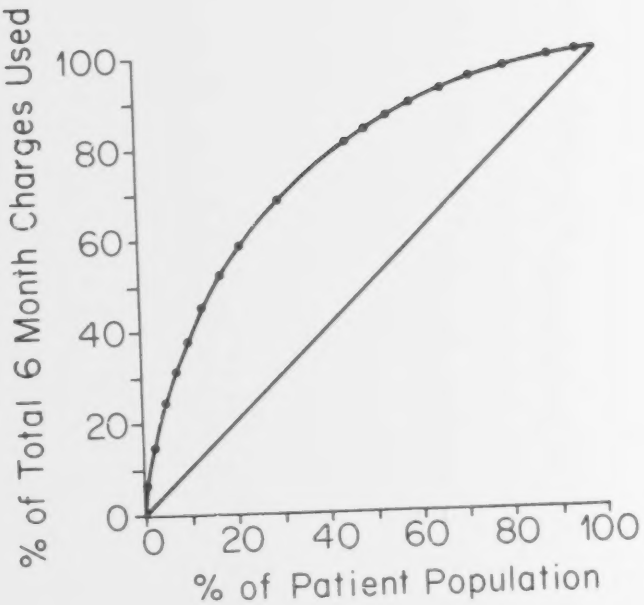


FIG. 4. Distribution of total six-month charges. Curved line represents actual distribution of charges; straight line indicates what distribution would have been had each patient incurred the same charges.

Discussion

Currently, very little is known about what kinds of patients are hospitalized in MICUs, and there is even less knowledge of the extent to which different institutions serve diverse MICU populations. How, for example, do teaching hospitals differ in this regard from community hospitals without house staff? Are there differences related to hospital or unit size? What are the urban/rural differences? Any attempt to develop guidelines for MICU use must be based on a thorough understanding of the differences in the populations served by these units.

One goal of the present study was to characterize a single hospital's medical intensive care population. The patients were found to be relatively young (mean age 54 years, only 19 per cent over 70 years of age). The diagnostic entities represented in the medical ICU population were spread over a wide range; only three homogeneous diagnostic groups each contained more than 5 per cent of the population. If this finding is confirmed in other institutions, the task of defining appropriate guidelines for MICU utilization will be that much more complex.

The present study identified a number of variables that appeared to be independently associated with survival: age, preadmission functional status, initial systolic blood pressure and diagnosis. This study also illustrates one particular problem with performing multivariate analysis on data from intensive care populations. Even with almost 500 study patients already screened to exclude surgical and CCU patients, controlling for more than three variables, particularly if one of them is diagnosis, partitions the group into very small subgroups. Such small sample sizes prevent even large differences in survival from attaining conventional levels of statistical significance. The findings of the study

TABLE 5. Characteristics of the High-Cost Subgroup

Characteristic	High-Cost Subgroup (%)	Whole Group (%)
Sex		
Male	53.1	45.1
Female	46.9	54.9
Mean age	53 years	54 years
Initial systolic blood pressure		
100 mmHg and less	12.5	13.8
Over 100 mmHg	87.5	86.2
Preadmission functional status		
Classes 0–1	12.9	16.0
Classes 2–3	45.2	42.2
Classes 4–5	42.0	41.8
Initial mental status (excluding overdoses)		
Alert	62.5	77.1
Abnormal	37.5	22.9
Length of ICU stay		
1–2 days	9.4	49.6
Over 10 days	56.2	10.3
Diagnosis		
Cardiovascular	25.0	22.5
Pulmonary	15.6	21.7
Gastrointestinal	15.6	14.9
Neurologic	6.2	7.4
Cancer	6.2	12.1
Chronic renal failure	12.5	3.3
Rheumatologic	12.5	3.9
Survival		
In-hospital	71.9	85.7
Six-month	53.1	71.6

with respect to survival should be confirmed in prospective studies at different institutions with larger numbers of patients. These further investigations will determine the extent to which the results of this study are generalizable and will permit a more detailed multivariate analysis of predictors of survival.

The finding that a simple functional status assessment yields important information about survival also requires confirmation in prospective studies. The retrospective use of this instrument in this study revealed that half of the patients perceived their functional status six months after discharge as equal to or better than their preadmission status. The direction of any bias that might have been introduced by the application of the instrument retrospectively is difficult to assess. On the one hand, relatives of nonsurvivors might tend to see the preadmission functional status of their relatives as worse than it actually was, because they remember only the severity of the illness. On the other hand, they might tend to recall the preadmission status as better than it actually was, because they remember the relatives as they were before the terminal illness. Similar arguments apply to patients recalling their own preadmission functional status. Prospective use of the instrument is the only way to measure the significance of these possibly confounding effects. If this meas-

ure is validated prospectively, it may prove a useful way to assess intensive care patient outcomes in addition to the more usual but cruder mortality measure.

There are relatively few data with which to compare the charge data gathered in this study. Other studies that have presented information on the costs of care for seriously ill patients analyzed very different kinds of patients in different time periods. After adjusting for inflation, the present investigation found costs that were roughly comparable to those measured by Cullen and his coworkers in their study of the most severely ill of their surgical ICU patients from 1972-73.[15] After a similar inflation adjustment, the present study also found costs roughly comparable to those measured by Griner for patients with pulmonary edema from 1970-71.[23] In their study of medical intensive care, Thibault and his colleagues measured hospital charges that were somewhat lower than those in this study for several diagnostic categories.[36] On the other hand, this study found costs substantially lower than those measured by Turnbull and his colleagues for cancer patients[35] and less than recent estimates of the cost of an institutionally based program of maintenance hemodialysis.[42] Unlike the present study, however, none of the studies previously cited separated ICU charges from ward charges. Comparisons were therefore made of total hospital charges, and no conclusions can be drawn regarding the relative charges generated by these different MICUs.

Until now much of the discussion of controlling ICU costs has focused on excluding patients with poor short-term prognoses from ICUs. A discussion of the limited conceptual and empirical evidence for the effectiveness of utilization review in containing hospital costs in general has been presented elsewhere[43]; it is equally applicable to this form of utilization review for intensive care patients. In the absence of explicit criteria for what is meant by poor short-term prognosis, physicians have been called upon to exercise their clinical judgment in deciding whom to admit to ICUs and whom to exclude. Rodman and his colleagues have shown that experienced critical care physicians were poorly able to distinguish eventual survivors from nonsurvivors in a mixed medical-surgical ICU, even at two and seven days after ICU admission.[44] This finding, if confirmed, suggests that more explicit criteria, based on empirical data, will be necessary if this effort is to be successful.

The present study sought to identify subgroups of medical intensive care patients with extremely low expected survival—categories of patients who might be candidates for exclusion from ICUs on the grounds of having poor short-term prognoses. The subgroup with the lowest six-month survival rate comprised cancer patients whose preadmission functional status placed them in classes 0–3; their 14 per cent survival rate is, however, clearly not low enough to justify systematic exclusion from ICUs.

In this study population, simple rules did not succeed in defining subgroups with extremely low survival rates. One cannot, for example, arbitrarily exclude all those over a certain age or all those with preexisting disabilities. Of the 15 study patients who were both older than 80 years of age and in the three lowest functional classes prior to their intensive care hospitalization, 12 survived to be discharged, and six survived six months.

One possible explanation of this finding is that the study hospital has already succeeded in excluding patients with very low expected survival from its medical ICUs. This was not, however, an explicit hospital policy at the time of the study. Decisions to admit or not to admit patients to intensive care were made by individual physicians and families. While it is likely that some critically ill patients were not admitted to intensive care as a result of this process,

this fact seems an unlikely explanation for the inability of this study to define with simple rules subgroups of patients with extremely low rates of survival. Moreover, one would expect that a tertiary hospital would have a relatively sicker intensive care population than other hospitals. Therefore, if one does not find patients with very low survival in a tertiary hospital, it may not be easier to do so in other, less specialized institutions.

Furthermore, each intensive care patient represents a virtually unique combination of primary and secondary disease processes, therapeutic regimens, and complications of therapy and disease. While in some sense this may be true for all patients, the significance of the differences among patients with similar diagnoses seems magnified in an intensive care population. This uniqueness, which is reflected in the diagnostic diversity described previously, makes the task of setting criteria for ICU exclusion even more difficult.

A final problem with putting this kind of control system into practice is the extreme difficulty encountered in obtaining agreement as to what level of expected survival rate would be low enough to justify exclusion from intensive care. Would survival have to be unprecedented, a situation that has been described for some badly burned patients?[45] Is 5 per cent low enough—a 1 in 20 chance? Furthermore, how should the views of patients and families be added into the balance? By now we are familiar with patients and families who decline heroic measures. But, conversely, should a patient's or family's desire for intensive care always supersede our criteria for exclusion, however they have been developed? The imponderability of these questions raises serious doubts about whether the effort to set criteria should be attempted at all.

This study demonstrated that, at least for medical intensive care, no single type of patient was responsible for a large portion of the total costs. Rather, the high-cost subgroup was spread among many ages, diagnostic groups and disability classes. This finding suggests that cost control cannot be achieved by focusing attention on any one homogeneous group of patients. Intensive care as a whole must be scrutinized more carefully.

Furthermore, the fact that more than half of the high-cost subgroup survived six months suggests that this investment of resources was not entirely without benefits. This finding highlights the very difficult choices that face intensive care physicians. Increasingly, we confront clinical situations in which the use of larger and larger amounts of medical resources may be associated with smaller and smaller additional benefits.

Measures that are more morally tenable exist for controlling ICU costs. First, one can examine a particular ICU population to ensure that no one has been admitted who is too healthy for intensive care. In the present study, for example, there was a subgroup of 16 patients among the 60 admitted for upper gastrointestinal bleeding who received no blood transfusions during their hospitalization; none of these patients died in the hospital, and only one died during the six-month follow-up period. Furthermore, of the 16 patients, 11 were admitted to the ICU with normal hematocrits, and seven left the hospital with normal hematocrits. In retrospect, it is reasonable to question their need for intensive care. Further investigation may lead to a clearer understanding of when or whether such patients require intensive care.

The second way in which physicians can influence ICU costs is through a careful attempt to reduce the intensity of intensive care. The present study measured average daily ICU charges of $1,120, a figure that includes $62 per day for inhalation therapy, a charge equal to the average daily

charge for all radiology services. Were all of these services necessary and effective? By how much could the $259 per day in laboratory charges be reduced if tests were ordered more prudently? These charges may overstate the opportunity for real cost savings, because charges for ancillary services may exceed their true costs. Substantial research will be required to determine which services are necessary and effective, and under what circumstances. The fruits of this effort, however, can result in better patient care at lower risk and lower cost.

The juxtaposition of highly visible and rapidly growing costs with less apparent and inadequately documented benefits can be expected to continue to attract public attention. It is incumbent upon those clinicians with an interest in effective intensive care to design and execute the research necessary to establish the place of intensive care in modern medicine. If this task is not addressed in a timely fashion, arbitrary restrictions on the use of intensive care issued by third-party payors may ensue.

Acknowledgments

The author is indebted to Linda Akita, R.N., Denice Allen, John Herman, Paul Hoffman, Susan Keel, R.N., J.D., Marion Winston, R.N., and Cheryl McCarthy, R.N., for their considerable assistance in data collection and processing; to Drs. John Beck, Robert Kane, Joseph Newhouse, Helen Smits and David Solomon for their insightful comments on earlier versions of this paper; to Marilyn Dean for her cogent suggestions on the final version; and to Dr. Robert Brook for his perceptive guidance throughout the study.

References

1. American Hospital Association. Hospital Statistics: 1971–1977 editions. Chicago, 1971–1977.

2. American Hospital Association. Hospitals 1959–1971, 33–45, Guide Issues.

3. Russell LB. The role of technology assessment in cost control. Presented at "Critical Issues in Medical Technology," conference sponsored by Harvard-MIT Division of Health Sciences and Technology, Boston, April 1980.

4. Griner P. Medical intensive care in the teaching hospital: Costs versus benefits. The need for an assessment Ann Intern Med 1973;78:581.

5. Ebert RE. Effectiveness and efficiency—How well are we doing? Am J Med 1979;66:191.

6. Martin SP, Donaldson MC, London CD, et al. Inputs into coronary care during 30 years. A cost effectiveness study. Ann Intern Med 1974;81:289.

7. Hofvendahl S. Influence of treatment in a coronary care unit on prognosis in acute myocardial infarction. Acta Med Scand 1973;519(suppl):1.

8. Hubner PJB, Goldberg MJ, Lawson CW. Value of routine cardiac monitoring in the management of acute myocardial infarction outside a coronary care unit. Br Med J 1969;1:815.

9. Reynell PC, Reynell MC. The cost-benefit analysis of a coronary care unit. Br Heart J 1972;34:897.

10. Stross JK, Willis PW III, Reynolds EW Jr. Effectiveness of coronary care units in small community hospitals. Ann Intern Med 1976;85:709.

11. Bloom BS, Peterson OL. End results, cost and productivity of coronary-care units. N Engl J Med 1973;288:72.

12. Pessi TV. Experiences gained in intensive care of surgical patients. A prospective clinical study of 1001 consecutively treated patients in a surgical intensive care unit. Ann Chir Gynaecol 1973; 62(suppl):1.

13. Davidson IA, Bargh W, Cruickshank AN, et al. Crush injuries of the chest. A follow-up study of patients treated in an artificial ventilation unit. Thorax 1969;24:563.

14. Morgan A, Daly C, Murawski BJ. Dollar and human costs of intensive care. J Surg Res 1973;14:441.

15. Cullen DJ, Ferrara LC, Briggs BA, et al. Survival, hospitalization charges and follow-up results in critically ill patients. N Engl J Med 1976;294:982.

16. Franco D, Bismuth H. Five years of experience with an intensive care unit specializing in abdominal surgery. Prog Surg 1977;15:37.

17. Renou P, Chang A, Anderson I, et al. Controlled trial of fetal intensive care. Am J Obstet Gynecol 1976;126:470.

18. Thompson T, Reynolds J. The results of intensive care therapy for neonates. J Perinat Med 1977;5:59.

19. Callahan JA, Spiekerman RE, Broadbent JC, et al. St. Mary's Hospital-Mayo Clinic medical intensive care units. II. Patient population. Mayo Clin Proc 1967;42:332.

20. Lepore MJ, Grace WJ. Role of the intensive care unit in gastroenterology 1969;51:493.

21. Clark TJH, Collins JV, Evans TR, et al. A review of experience operating a general medical intensive care unit. Br Med J 1971;1:158.

22. Spagnolo SV, Hershberg PI, Zimmerman HJ. Medical intensive care unit: Mortality rate experience in large teaching hospital. NY State J Med 1973;73:754.

23. Griner P. Treatment of acute pulmonary edema: Conventional or intensive care? Ann Intern Med 1972;77:501.

24. Rogers RM, Weiler C, Ruppenthal B. Impact of the respiratory intensive care unit on survival of patients with acute respiratory failure. Chest 1972;62:94.

25. Piper KW, Griner PF. Suicide attempts with drug overdose; outcomes of intensive vs. conventional floor care. Arch Intern Med 1974;134:703.

26. Askrog VF, Dam WH, Hagelsten JO. Treatment of patients with pulmonary oedema in an intensive therapy department. Br J Anaesth 1967;39:252.

27. Sluiter HJ, Blokzijl EJ, van Dijl W, et al. Conservative and respirator therapy of acute respiratory insufficiency in patients with chronic obstructive lung disease. Am Rev Respir Dis 1972;105:932.

28. Jessen O, Kristensen HS, Rasmussen K. Tracheostomy and artificial ventilation in chronic lung disease. Lancet 1967;ii:9.

29. Bendixen HH. The cost of intensive care. In: Bunker JP, Barnes BA, Mosteller F, eds. Costs, Risks, and Benefits of Surgery. New York: Oxford University Press, 1977.

30. Linton RC, Walker FW, Spoerel WE. Respirator care in a general hospital: A five-year survey. Can Anaesth Soc J 1965;12:450.

31. Campbell D, Reid JM, Telfer ABM, et al. Four years of respiratory intensive care. Br Med J 1967;4:255.

32. Petty TL, Lakshminarayan S, Sahn SA, et al. Intensive respiratory care unit. Review of ten years' experience. JAMA 1975;233:34.

33. Soler NG, Bennett MA, FitzGerald MG, et al. Intensive care in the management of diabetic ketoacidosis. Lancet 1973;i:951.

34. Nunn JF, Milledge JS, Singaraya J. Survival of patients ventilated in an intensive therapy unit. Br Med J 1979;1:1525.

35. Turnbull AD, Carlon G, Baron R, et al. The inverse relationship between cost and survival in the critically ill cancer patient. Crit Care Med 1979;7:20.

36. Thibault GE, Mulley AG, Barnett GO, et al. Medical intensive care: Indications, interventions, and outcomes. N Engl J Med 1980;302:938.

37. Griner PF, Liptzin B. Use of the laboratory in a teaching hospital. Implications for patient care, education, and hospital costs. Ann Intern Med 1971;75:157.

38. Schroeder SA, Showstack JA, Roberts HE. Frequency and clinical description of high-cost patients in 17 acute-care hospitals. N Engl J Med 1979; 300:1306.

39. Zook CJ, Moore FD. High-cost users of medical care. N Engl J Med 1980;302:996.

40. A report of the Clinical Care Committee of the Massachusetts General Hospital: Optimum care for hopelessly ill patients. N Engl J Med 1976;295:362.

41. Stewart AL, Ware JE Jr, Brook RH, et al. Conceptualization and Measurement of Health for Adults in the Health Insurance Study. Volume II, Physical Health in Terms of Functioning. R-1987/2-HEW. Santa Monica: The Rand Corporation, July 1978.

42. Stange PV, Sumner AT. Predicting treatment costs and life expectancy for end-stage renal disease. N Engl J Med 1978;298:372.

43. Chassin MR. The containment of hospital costs: A strategic assessment. Med Care 1978;16(suppl): October: 1.

44. Rodman GH, Etling T, Civetta JM, et al. How accurate is clinical judgment? Crit Care Med 1978;6:127 (abstract).

45. Imbus SH, Zawacki BE. Autonomy for burned patients when survival is unprecedented. N Engl J Med 1977;297:308.

MEDICAL CARE
February 1982, Vol. XX, No. 2

Predicting the Outcome of Primary Care

DIANA DRYER WRIGHT, M.D.,* AND ROBERT L. KANE, M.D.†

Each physician's ability to treat disease is limited by his/her ability to discriminate among patients on the basis of risk. The relationships of physician expectation of outcome (prognosis) to actual outcome and to cost were determined for 1,757 patients seeking primary care. Outcome was measured by a seven-level functional-status scale; patients who returned to their usual level of function after an acute illness were defined as having good outcomes. Although 24 per cent of patients experienced bad outcomes, physicians had anticipated only 6 per cent. Physician's predictions of bad outcomes had a sensitivity of 13.6 per cent and a specificity of 96.9 per cent. Patients with bad outcomes had slightly higher laboratory costs than did patients with good outcomes, but a much larger increase was seen in laboratory, office and total costs for all patients for whom physicians expected bad outcomes, regardless of the actual results. A feedback loop is recommended to provide a better perspective on outcome and eventually to improve the efficiency and cost benefit of the medical decision-making process.

THE INCREASING COST of medical care has made primary-care physicians aware that, as an integral part of any health care system, they must increase their efforts for more efficient, as well as more effective, treatment.[1,2] Both their patients and those responsible for third-party payment are demanding answers about what they are buying and how much they should be paying for it.[3] Physicians are exploring how current methods of assessing quality of care are related to cost.[4,5]

Primary-care physicians face a special challenge because they have been identified as the entry point into the health care system. In a context of high volume, the physician must discriminate between substantive and self-limited illness in order to affect outcome.[6] Traditional clinical judgment has been shown to be relatively imprecise, even in such defined clinical entities as systemic lupus erythematosis[7] and peptic ulcer disease.[8] Increased reliance on laboratory tests is not necessarily related to improvement in outcome.[9–12]

Far less information is available about the course of the cluster of complaints that represent the bulk of primary care; the typical diagnosis-oriented approach may not be useful in evaluating the course of mild, nonspecific or self-limited illnesses. The assumption is that a diagnosis implies a prognosis,[13] but too often diagnostic deci-

* Department of Family and Community Medicine, University of Utah College of Medicine.

† School of Medicine, UCLA, and The Rand Corporation, Santa Monica, California.

Work for this study was supported by Public Health Service Award No. 1 RO1 HSO 1596 from the National Center for Health Services Research.

Reprint requests: Robert L. Kane, M.D., The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406.

0025-7079/82/0200/0180/$00.90 © J. B. Lippincott Co.

sions must be made on the basis of inadequate data by physicians who have not been trained to deal with uncertainty. The primary-care physician's ability to predict accurately patient outcomes is therefore a subject of importance[14] because the physician's expectation of outcome can influence the application and success of treatment.

Method

To study the relationships between prognosis and outcome and between prognosis and cost in primary care, data on prognosis, outcome and cost were collected on all patients who sought care for any acute complaint during a nine-month period at two model family practice clinics at the University of Utah.[15] Less than 1 per cent of the patients refused to participate. Patients who required hospitalization (0.5 per cent), patients who could not be reached for follow-up (8 per cent) and patients who experienced a second, separate episode of illness before follow-up (6 per cent) were excluded from the final sam0le of 1,757 treatment episodes. Return visits for the same problem were included as part of a single episode. This study thus focused on illnesses that had produced a change in function and offered an opportunity to evaluate the consequences of the physician's intervention.

The physicians included first-, second-, and third-year family practice residents and family practice faculty; physician's assistants also saw some patients. All physicians were informed in advance that the study would focus on patient outcome, and their cooperation was solicited. No attempt was made by the investigators to influence physicians' treatment behavior.

Before each patient was seen by the physician, a trained interviewer obtained information about the patient's functional status during the six months prior to the clinic visit and his/her current functional status, reflecting the degree of impairment imposed by the illness. Functional-status ratings were assigned by the interviewer on the basis of the activities reported by the patient. Williamson's seven-level scale was used because it is sensitive to relative function (i.e., it discriminates appropriately for the individual's age and occupation).[16] The categories were: 1) full activity without symptoms; 2) an underlying phys-

TABLE 1. Physician Prediction of Outcome by Actual Outcome

Physician Prediction of Outcome	Actual Outcome		
	Bad	Good	Total
Poor prognosis	56	42	98
Good prognosis	357	1302	1659
Total	413	1344	1757

Sensitivity of prediction = $\frac{56}{413} = 0.136.$

Specificity of prediction = $\frac{1302}{1344} = 0.969.$

Predictive value of a positive test (poor prognosis) = $\frac{1320}{1659} = 0.571.$

Predictive value of a negative test (good prognosis) = $\frac{56}{98} = 0.785.$

ical or laboratory abnormality without symptoms; 3) symptoms, but with full activity; 4) symptoms with restricted activity; 5) limited mobility (i.e., housebound); 6) confinement to bed; and 7) death.

After a patient was seen, the physician was asked to predict 1) the length of time necessary for the patient to receive the maximum benefit from treatment and 2) the functional status the patient should achieve by that time. Any patient whom the physician did not expect to return to his/her usual functional status was defined as having a poor prognosis. The physician was allowed to delay response for up to one week after the patient's first visit in order to make the prognosis on the basis of as much information as possible.

At the time specified by the physician as appropriate for follow-up (not less than two weeks or longer than six months after the visit), the interviewer visited each patient at home to determine his/her actual functional status as a consequence of treatment. The interviewer also obtained any laboratory specimens needed to evaluate the effectiveness of treatment. A good outcome was defined as one in which the patient's functional status on follow-up was equal to or better than his/her usual functional status prior to illness. The number of patients for whom the physicians reported the discovery and persistence of previously unidentified physical or laboratory abnormalities that would prevent a patient from returning to a completely normal functioning status (level 1) was very small (less than 3 per cent).

Physician charges for the office visit and laboratory and x-ray fees were obtained from clinic and hospital billing files. All charges recorded between the initial clinic visit and the follow-up date were included. The physician fee schedule was competitive with that of local private physicians and did not vary according to level of training. Medication costs were determined by the prices charged in the adjoining hospital pharmacy for all medications prescribed by physicians, whether or not the prescriptions were filled. "Other" costs (for services such as physical therapy) were rare; when present, they were included in the total cost and were not examined separately.

Results

The percentage of patients identified by their physicians as having a poor prognosis was very small (6 per cent), as shown in Table 1. In fact, 24 per cent of patients experienced a bad outcome. Although the fraction of patients with a bad outcome was much higher among poor-prognosis patients (56/98), 357 patients with bad outcomes were among the 1,659 patients for whom the physicians expected a good outcome.

These findings can be described in terms of clinical judgment as the predictive value of a positive test (here the proportion of poor-prognosis patients who actually experienced bad outcomes) and the predictive value of a negative test (the proportion of good-prognosis patients who experienced good outcomes), which are here 57 per cent and 79 per cent, respectively. A high predictive value of a positive test is of particular importance in any large-volume operation, such as a mass screening program or primary care itself, in order to ensure that a large number of healthy people are not falsely labeled as ill and subjected to unnecessarily invasive or expensive further testing or treatment. Conversely, a high predictive value of a negative test should protect individuals with true disease from escaping detection and missing appropriate intervention.

Two other parameters frequently mentioned as measures of the usefulness of a screening test or a discriminating variable are the sensitivity and specificity of detecting the disease.[17] In this instance, sensitivity refers to the proportion of bad outcomes that were correctly predicted, and

specificity refers to the proportion of good outcomes correctly predicted. The data in Table 1 reveal a sensitivity of 13.6 per cent and a specificity of 96.9 per cent. The crucial difference between the sensitivity and specificity on the one hand, and the predictive value of a positive test and the predictive value of a negative test on the other, is that the first pair refers to the effectiveness or accuracy of a single test in detecting disease and the second pair refers to the consequences of that detection for patients. Both kinds of information about physicians' judgments are important, and the emphasis that may be placed on one or the other relates to the prevalence of the condition under consideration and the costs of applying or missing diagnosis and treatment. In this study we examined the occurrence of a condition (bad outcomes) that was thought to be rare but, in fact, was very common. Therefore, the more frequent error that physicians made was failure to detect patients who would experience bad outcomes.

In order to evaluate the importance of the large number of patients with bad outcomes that were not recognized by their physicians, one can consider how the predictive value of positive and negative tests and the sensitivity and specificity differed by severity of illness, by diagnosis and by the cost of management. Table 2 presents these differences by frequently occurring specific complaints. Problems that a physician might expect to be self-limited (e.g., pharyngitis, upper respiratory infection, flu) had low sensitivities and high specificities, very low predictive value of positive tests and reasonably high predictive value of negative tests. Sensitivity was greater for the more serious problems but was in no case above 40 per cent; however, the predictive value of a positive test increased remarkably for the more serious problems.

Table 3 shows that none of the four measures varied noticeably according to

TABLE 2. Prediction of Outcome for Common Problems

Problem	Sensitivity (Proportion of Bad Outcomes Correctly Predicted)	Specificity (Proportion of Good Outcomes Correctly Predicted)	Predictive Value of a Positive Test (Proportion of Poor Prognoses with Bad Outcomes)	Predictive Value of a Negative Test (Proportion of Good Prognoses with Good Outcomes)	Per Cent of Patients with Bad Outcomes	Number of Patients
Pharyngitis	0.000	0.981	0.000	0.825	17.2	192
Upper respiratory infection	0.000	0.986	0.000	0.802	19.5	261
Otitis media	0.038	0.953	0.167	0.803	19.5	133
Flu	0.000	1.000	—	0.796	20.4	49
Dermatitis	0.000	0.895	0	0.708	26.9	52
Cough and bronchitis	0.074	0.971	0.500	0.725	28.4	95
Neck and back pain	0.375	0.919	0.667	0.773	30.2	53
Abdominal pain	0.258	0.982	0.889	0.701	36.0	86
Headache and malaise	0.267	0.938	0.800	0.577	48.4	62

TABLE 3. Prediction of Outcome by Patient Status at Time of First Visit

Patient Presenting Status	Sensitivity (Proportion of Bad Outcomes Correctly Predicted)	Specificity (Proportion of Good Outcomes Correctly Predicted)	Predictive Value of a Positive Test (Proportion of Poor Prognoses with Bad Outcomes)	Predictive Value of a Negative Test (Proportion of Good Prognoses with Good Outcomes)	Per Cent of Patients with Bad Outcomes	Number of Patients
Asymptomatic but abnormal physical or laboratory examination	0.364	0.941	0.667	0.821	24.4	45
Symptomatic but no limitation of activity	0.134	0.972	0.611	0.773	24.8	993
Symptomatic with minor limitation of activity	0.110	0.969	0.484	0.803	21.1	646
Symptomatic with major limitation of activity	0.200	0.943	0.571	0.758	27.4	73

TABLE 4. Prediction of Outcome by Type of Provider

Type of Provider	Sensitivity (Proportion of Bad Outcomes Correctly Predicted)	Specificity (Proportion of Good Outcomes Correctly Predicted)	Predictive Value of a Positive Test (Proportion of Poor Prognoses with Bad Outcomes)	Predictive Value of a Negative Test (Proportion of Good Prognoses with Good Outcomes)	Per Cent of Patients with Bad Outcomes	Number of Patients
First-year resident	0.090	0.978	0.546	0.785	22.7	295
Second-year resident	0.076	0.963	0.387	0.775	23.3	674
Third-year resident	0.210	0.957	0.618	0.786	24.8	403
Faculty	0.197	0.982	0.824	0.743	29.7	239
Physician's assistant	0.167	0.984	0.600	0.894	12.3	146

the patient's presenting status when ill. Among types of providers (Table 4), third-year residents and faculty demonstrated higher sensitivities and higher predictive value for positive tests than did first- or second-year residents. Physician's assistants predicted bad outcomes more accurately than did first- or second-year residents, but less well than did third-year residents and faculty; this might reflect the influence of supervisory faculty. These differences were small but provide some evidence for a trend toward improvement in accuracy of prognosis with increasing level of training and/or experience.

The long-term economic impact of an unrecognized bad outcome in primary care is not known, but these data can provide some perspective on the cost of a poor prognosis. Differences in cost between the good-prognosis and poor-prognosis patients and between the patients who experienced good and bad outcomes are shown in Tables 5–8. A poor prognosis was associated with higher total costs and increased laboratory and office visit fees. Because a fixed amount ($12) was charged for each office visit, physician office costs indicate the number of patient visits. The dollar differences therefore represent much more laboratory work but only a few more return visits for those patients whose physicians believed that they would not regain their usual functional status. Bad outcomes were also associated with higher laboratory and total costs, but the dollar differences between patients with good and bad outcomes were much less than were those between poor-prognosis and good-prognosis patients. Among patients with a good prognosis, dollar differences in costs between patients with good and bad outcomes were not striking. However, among patients with a poor prognosis, even those who did experience good outcomes had higher costs than did any of the good-prognosis patients, and the increases in total and laboratory costs for poor-prognosis patients with bad outcomes were remarkable. The data thus suggest that prognosis has a much greater effect on the immediate cost of primary care than does outcome.

TABLE 5. Average Costs by Prognosis and by Outcome: Total Costs

Prognosis	Actual Outcome		
	Bad ($)	Good ($)	Overall ($)
Poor	33.05	26.48	30.23
Good	21.36	19.83	20.16
Overall	22.95	20.04	20.72

Discussion

Because a physician's expectation of outcome appears to be an important determinant of cost, it is essential for physicians to become more aware of the factors that influence the cost-generating decisions that are made in the course of a patient's illness. For example, on a patient's first visit, the physician can choose one or a combination of several options: 1) treat the patient on the basis of available information; 2) request additional information, usually in the form of laboratory studies; 3) prolong the period of observation, with or without the initiation of treatment or the gathering of new data, by scheduling a follow-up visit; 4) do nothing more than reassure the patient.

Each of these choices will affect cost and may affect outcome; they must be

TABLE 6. Average Costs by Prognosis and by Outcome: Office Costs

Prognosis	Actual Outcome		
	Bad ($)	Good ($)	Overall ($)
Poor	15.79	13.67	14.88
Good	12.77	12.59	12.63
Overall	13.18	12.62	12.75

TABLE 7. Average Costs by Prognosis and by Outcome: Laboratory Costs

Prognosis	Actual Outcome		
	Bad ($)	Good ($)	Overall ($)
Poor	14.09	7.82	11.42
Good	5.25	3.72	4.05
Overall	6.46	3.85	4.46

evaluated individually in terms of the physician's awareness of risk and benefit to each patient. Unfortunately, our data indicate that primary-care physicians' prognoses may both be incorrect for a large proportion of patients whom they identify as having a poor prognosis (shown by the low predictive value of a positive test) and may also miss a substantial number who fail to return to their usual status after an acute illness (low predictive value of a negative test). Some physicians might comfort themselves that such patients will return voluntarily for further treatment,[18] but we do not know the consequences—physical, psychological or economic—of the intermediate delay.

The fact that physicians did order extra laboratory tests, even during the first visit, for patients with bad outcomes suggests that they may have been uncomfortable about the nature or projected course of certain illnesses in spite of the fact that they predicted good results. However, because a physician in our study could delay making a prognosis until after he/she had received the laboratory reports, the test results must not have been sufficiently indicative of trouble to prompt him/her either to request a follow-up visit or to lower expectation of outcome. Without an interviewer to obtain follow-up data, and with no return visit, the physician might never learn about the patient's remaining impairment. The physician would thus have no regular means of determining whether the investment in the cost of extra laboratory work had produced any benefit. We need much more data before we will know if the additional cost of early identification and treatment of a larger proportion of the primary-care patients who experience bad outcomes offsets the cost of those bad outcomes exerting their effects on function over time, causing patients to seek subsequent care with the same probability of incorrect prognosis.

TABLE 8. Average Costs by Prognosis and by Outcome: Medication Costs

Prognosis	Actual Outcomes		
	Bad ($)	Good ($)	Overall ($)
Poor	3.18	4.62	3.80
Good	3.24	3.17	3.18
Overall	3.23	3.21	3.22

To enable each physician to examine the consequences of his/her actions, we believe that two affirmative steps would be valuable. First, each physician could document the prognosis at each patient visit. Second, each physician could develop an organized system of follow-up, perhaps using inexpensive telephone contacts[19] in order to verify and improve accuracy. This exercise should not only improve the quality of care we give patients, but should also demonstrate our efforts to control costs on an individual basis.

Conclusion

In summary, primary-care physicians' prognostic abilities are imprecise, with low sensitivity and low predictive values for both positive test (poor prognosis) and negative test (good prognosis). During initial contacts with patients, physicians apparently order laboratory tests that they do not pursue with follow-up and do not use to adjust their expectation of patient outcome. The additional investment, therefore, does not seem to assist them in identifying a

large number of patients with bad outcomes. A feedback loop is recommended to provide a better perspective on outcome and eventually to improve the efficiency and cost benefit of the medical decision-making process.

References

1. Draper P, Smits HL. The primary-care practitioner: Specialist or jack-of-all-trades. N Engl J Med 1975;293:903.

2. Garfield SR, Collen MF, Feldman R, et al. Evaluation of an ambulatory medical-care delivery system. N Engl J Med 1976;294:426.

3. Eisenberg JM, Rosoff AJ. Physician responsibility for the cost of unnecessary medical services. N Engl J Med 1978;299:76.

4. Wright DD, Kane RL, Snell GF, et al. Costs and outcomes for different primary care providers. JAMA 1977;238:46.

5. Pauker SG, Kassirer JP. Therapeutic decision making: A cost benefit analysis. N Engl J Med 1975;293:229.

6. Schwartz WB, Gorry GA, Kassirer JP, et al. Decision analysis and clinical judgment. Am J Med 1973;55:459.

7. Shapiro AR. The evaluation of clinical predictions. N Engl J Med 1977;296:1509.

8. Peterson WL, Sturdevant RAL, Frank HS, et al. Healing of duodenal ulcer with antacid regimen. N Engl J Med 1977;297:341.

9. Gorry GA, Pauker SG, Schwartz WB. The diagnostic importance of the normal finding. N Engl J Med 1978;298:486.

10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926.

11. Daniels M, Schroeder SA. Variation among physicians in use of laboratory tests. II. Relation to clinical productivity and outcomes of care. Med Care 1977;15:482.

12. Sturdevant RAL, Stern D. Accuracy of physicians' prediction of cholecystography results. Med Care 1977;15:488.

13. Mushlin AI, Appel F, Pozen JT, et al. The use of patient-reported problem status in evaluating ambulatory care. Clin Res 1975;23:267.

14. Koran LM. The reliability of clinical methods, data and judgments, Parts I and II. N Engl J Med 1975;293:642.

15. Kane RL, Gardner J, Wright DD, et al. A method for assessing the outcome of acute primary care. J Fam Pract 1977;4:1119.

16. Williamson JW. Outcomes of health care: Key to health improvement. In: Hopkins CE. Methodology of identifying, measuring and evaluating outcomes of health service programs, systems and subsystems. Washington, D.C.: Government Printing Office, 1977.

17. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975;293:211.

18. Mushlin AI, Appel FA. Diagnosing potential noncompliance: Physicians' ability in a behavioral dimension of medical care. Arch Intern Med 1977;137:318.

19. Henk M, Froom J. Outreach by primary-care physicians. JAMA 1975;233:256.

MEDICAL CARE
February 1982, Vol. XX, No. 2

Hospital Discharge Queues in Massachusetts

LEONARD W. GRUENBERG, PH.D., AND THOMAS R. WILLEMAIN, PH.D.

There is growing concern over the inappropriate utilization of health care facilities. The high cost of hospital care and the apparent shortage of nursing home beds have focused attention on one aspect of this problem: the clinically unnecessary days (sometimes called "administrative days" or "ADs") spent in hospitals by patients who are awaiting placement in long-term care facilities. In this study, data from a 1976 Massachusetts Department of Public Health survey of patients backed up in hospitals were analyzed to determine the magnitude of the problem and to examine the influence of several major factors that had been hypothesized in previous studies to contribute to the backup. We demonstrate that the average delay of a patient found waiting in a "snapshot" survey (which is often used to estimate the magnitude of the problem) is significantly greater than the average delay experienced by a typical discharged patient. We show that there are at least two major factors that influence the delay time: nursing home preferences in accepting certain types of patients and nursing home occupancy rates in the hospital service area. Neither medical-surgical occupancy rate nor the number of AD patients waiting in the hospital was significantly correlated with the delay time.

SPIRALING HEALTH CARE COSTS have sparked growing concern over the inappropriate utilization of health care facilities. The high cost of hospital care, as well as the apparent simplicity of a number of proposed remedies, has focused attention on one aspect of this problem—an aspect that is known as the problem of "administrative days."

Administrative days are a phenomenon of the juncture between acute and long-term care. Acute hospital patients with chronic disease or disability who require less intense but more prolonged treatment sometimes find their hospital discharges delayed because a suitable placement has not been arranged. The days spent in hospitals merely awaiting placement are referred to as "administrative days" (ADs) to distinguish them from that portion of the length of stay deemed clinically appropriate.

The purpose of this article is to examine the extent to which certain factors appear to influence the length of time it takes for individuals in hospitals to be placed in long-term care facilities. The factors that we have studied as possible determinants of the delay—patient characteristics, hospital occupancy rate and area wide nursing home occupancy rates—were chosen because they relate directly to solutions that have been proposed to minimize the undesirable effects of the AD problem.

From the University Health Policy Consortium, Brandeis University, Waltham, Massachusetts.

Supported in part by a Health Care Financing Administration grant to the University Health Policy Consortium, #18-P-97038/1-03.

Reprint requests: Leonard Gruenberg, UHPC, Heller School, Brandeis University, Waltham, MA 02254.

0025-7079/82/0200/0188/$01.20 © J. B. Lippincott Co.

Administrative days may be perceived either as a problem in themselves or as a symptom of other problems. Those concerned about the level of hospital expenditures object to ADs on grounds of cost. Those concerned about overmedicalization of long-term care object to the extension of hospital stay on grounds of quality of care and quality of life. Others see in ADs a reflection of some other health planning issue: nursing home bed shortages (which might create backlogs), low hospital occupancy rates (which might tempt hospitals to prolong stays), or discrimination against certain classes of patients (which might be interpreted as either venal or prudent, depending on further analysis of causes).

Whether regarded primarily as a problem in themselves or as evidence of more basic difficulties, ADs have been attracting the attention of both the acute and long-term care communities for some time, and interest is growing. In Massachusetts the state chapter of the Society for Hospital Social Work Directors of the American Hospital Association conducted two studies (1972 and 1973); the Massachusetts Hospital Association conducted three studies (1974, 1976 and 1980); the state Department of Public Health conducted three (1974, 1976 and 1980); and the state Department of Public Welfare produced two studies in 1980 alone. In New York, the state Council of Professional Standards Review Organizations undertook a study of the AD problem in 1979 and followed with another in 1980, labeling the problem a "crisis." Shapiro, Ross and Kavanaugh[1] documented the extent of the problem in Manitoba, Canada. The Office of the Inspector General, Department of Health and Human Services[2] studied the AD problem in several states at the request of the Secretary and predicted that it would grow more serious.

Scanning these studies, one discovers two sets of issues. The first concerns the estimation of the *magnitude* of the AD problem; the second concerns its *explanation*. Every AD study we have read has been based primarily on a one-day "snapshot" survey designed to estimate the magnitude of the problem at a particular time. Such snapshot surveys introduce distortions that appear not to be widely appreciated, so we will discuss the issue of distortion. Our main task, however, is to help with the issue of explanation by testing three different interpretations of the AD phenomenon against data from a study conducted in 1976 by the Massachusetts Department of Public Health.

The three competing explanations have been distilled from the excellent summary paper by Beatrice.[3] One view holds that ADs arise because nursing homes discriminate against certain classes of patients, being reluctant to admit those whose care will be paid for by Medicaid and especially reluctant to admit Medicaid patients with "heavy care" needs and/or behavior problems. A second view holds that ADs arise because there are too few nursing home beds and/or home care services to accommodate the growing number of frail elderly, thereby forcing backlogs in the hospitals. The third view holds that hospitals are responsible for ADs, either because AD patients represent "easy money" when the hospital is faced with a low occupancy rate or because the hospital discharge process is inefficient.

All three theories have received some support in the literature. The 1979 Massachusetts Hospital Association study[4] concluded that the supply of nursing home beds was too low in general and that nursing homes were reluctant to admit Medicaid and/or "heavy care" patients. As might be expected, the Hospital Association report stressed the diligence of hospitals in seeking placement. On the other hand, the 1980 Massachusetts Department of Public Welfare study[5] declared that patient characteristics played a very minor

role in the AD problem as compared with disinterested or incompetent placement efforts by hospitals. The Department's study agreed that nursing home bed shortages exacerbated the problem but also noted patterns of "sub-area politics" that restricted the effective bed supply to a number below that nominally available. Nationally, the Inspector General's report[2] cited evidence for all three theories: nursing home selectivity, hospital incentives to keep AD patients and nursing home bed shortages. In this study we first take a fresh look at the magnitude of the AD problem, then examine the evidence for the three competing explanations. We also draw inferences about useful responses.

Data

The data set used to study the magnitude and determinants of the AD problem is a survey conducted by the Massachusetts Department of Public Health (MDPH) on March 10, 1976 and followed up on April 19, 1976.[6] The survey was a repeat of an earlier one carried out in 1974.[7] A copy of the survey instrument is available from the authors. Respondents were discharge planners in each of the acute and chronic disease/rehabilitation hospitals in Massachusetts. Each was asked to report on those patients who were awaiting discharge to a lower level of care on the morning of the survey. The follow-up survey six weeks later was to ascertain the disposition of each case. A total of 1,322 individuals were found to be awaiting placement on March 10, 1976. Of these, we will focus only on the 620* adults who were waiting in acute hospitals for placement in a skilled nursing facility (SNF) or an intermediate care facility (ICF), since it is the acute hospital/nursing home interface that is the subject of this study.

The survey provided two different measures of ADs for each patient. One was the time the patient had spent on AD status up to the day of the survey. The other was the time from the survey date until the patient left the AD queue, either because she was placed, died, or again required the clinical services of the hospital; for the 12% of our patients whose delay exceeded the six-week follow-up period, the exact duration could not be recorded. (We conservatively recorded these stays as 43 days.) We took as the estimate of ADs for each patient the sum of the periods before and after the survey date.

The survey recorded roughly 30 patient descriptors, including demographics, measures of dependence in activities of daily living (ADL), rehabilitation potential, behavior problems, types of service required, level of nursing home care needed and sources of payment for care. The survey also recorded the identification number and health systems agency of the hospital in which the patient waited. These identifiers allowed linkage to other data bases on hospital and nursing home occupancy.

Estimation of ADs

The procedure usually used to estimate the magnitude of the AD problem has an obvious advantage and a less well-known disadvantage. That procedure is to obtain a "snapshot" of the AD problem on a particular day. A snapshot survey is economical and practical, since effort—and compliance—need only be mustered for a single day's data collection. To the extent that the chosen day is typical, the snapshot provides a good estimate of the number of patients awaiting placement.

The disadvantage of a snapshot survey is that if the data are used to estimate the delay of a "typical" patient who becomes

* Of these 620 patients, 45 lacked a full set of data on personal characteristics. These missing cases had very long waiting times (mean = 156 days, SD = 347 days). Most of the analyses focus on the 575 patients for whom a full description was available.

ready for discharge on a given day, the resulting estimate is severely biased: the average delay of persons found in the snapshot is significantly greater than the average delay of a typical person being discharged. The bias arises because patients with longer delays are more likely to be found waiting on any particular day than are patients with brief delays. Many of the latter can come and go in the system without detection by a snapshot. In contrast, patients with long delays are at proportionately greater risk of discovery. Of course, an additional problem with snapshots is that only that part of the total delay that has already transpired can be recorded; a follow-up survey is needed to determine the total delay for each patient.

When certain mathematical conditions are satisfied, the theory of "length-biased sampling"[8] or "random incidence"[9] can be used to convert the data on partial delays found in a snapshot into unbiased estimates of total delay as found in a longitudinal study. One key result of this theory is that the mean interval measured in a snapshot survey is a multiple of the true mean number of ADs. The multiplier involves the degree of variability in the distribution of ADs: the more variability, the greater the overestimate obtained from a snapshot. It can even happen that one group of patients with a shorter mean placement delay than a second group can have a longer mean delay in a snapshot; this would arise if the first group contained a significant minority of individuals with very long delays and the second group did not.

That the 1976 MDPH survey included a follow-up after six weeks allows a demonstration of the magnitude of the overestimate of placement delay produced by a snapshot. Since even the six-week follow-up period was not long enough to capture the full delay of many patients, estimates of mean delay will be distorted; we will therefore cite median delays instead. Consider adult, acute hospital patients awaiting placement in a nursing home. Of 620 such patients, 102 became ready for discharge on the day of the snapshot survey ("new patients"), and 518 were found already waiting ("old patients"). The median total wait for the old patients was 21 days, whereas the median for new patients was only five days. This fourfold difference should be significant for discussions of the magnitude and nature of the AD problem. As we note below, the same phenomenon appeared in the utilization of nursing homes; according to results obtained from the 1977 National Nursing Home Survey,[10] the median length of stay was less than three months among nursing home discharges but greater than 12 months among nursing home residents.

Having noted these distortions in estimates of delay, we are nonetheless forced to use the survey figures in our subsequent analyses. Since these analyses concern comparative rather than absolute magnitudes, the data suffice for our purposes. The reader is cautioned to remember, though, that the mean delays we cite are inflated by the length-biased sampling of the survey.

Explanation of ADs

We gained considerable insight into the waiting time problem by examining the distribution of waiting times. Table 1 shows how the total number of AD patients from the entire group ("new" and "old" patients) who were found to be still awaiting placement decreased with time after the date of the survey. It can be seen that the number waiting drops sharply during the first few days after the survey: in fact, of the 620 patients who were found to be waiting on the day of the survey, 156 (slightly more than one quarter) were placed within two days. On the other hand, the distribution has a long "tail." Observe in Table 1, for example, that one quarter of the patients were still waiting after three

TABLE 1. Per Cent of Patients Remaining in Discharge Queue after Survey Date

Weeks after Survey	Per Cent Remaining
1	51.9
2	34.7
3	25.5
4	17.9
5	13.9
6	11.6

weeks and 12 per cent were still not placed when the follow-up was made six weeks after the survey date.

This analysis suggests that there are (at least) two types of patients: those "easy to place" and those "hard to place." However, the considerable spread in the distribution points to the fact that these two groups may not be very well defined, i.e., that some patients who look quite similar to the patients in the hard-to-place group may have been placed in a short period of time and vice versa.

To proceed with the analysis, half of the data (every second case) was first set aside to serve as a confirmatory data set. The remaining cases served as an exploratory [illegible] a set on which to develop hypotheses [illegible]odels.

Patient Characteristics

The first step in the analysis was to look at two groups of patients who represented extremely short and extremely long stays. Included were 28 patients whose total waiting time (i.e., the time interval between when they became ready and when they were placed) was two days or less and 115 whose wait exceeded 30 days. We found major differences between the two groups on several variables. Patients in the "easy" or short-stay group were found to be more likely to need physical therapy, to be rehabilitatable and likely to return to the community, to have a payment source other than Medicaid, and to be admitted to the hospital from home or from a nursing home (as opposed to from some other institutional or group setting). Patients in the long-stay group were more likely to be totally dependent if they were to return home; their household was most often regarded as unable to cope physically and emotionally with the patient, and they were more likely to have behavior problems. This analysis of cases with extremely long or short stays, together with other multivariate analysis and anecdotal descriptions of the system provided by Welfare Department officials and discharge planners, led us to examine groupings or profiles of patients using the following variables as the building blocks of the profiles:

1. Source of Admission to the Hospital—Patients were grouped into three subgroupings: admitted to the hospital from home, from a nursing home, or from some other institutional or group setting (labeled as "other").

2. Likely Source of Payment in the Nursing Home—Patients were classified as either Medicaid or non-Medicaid (the latter being mostly Medicare or self-pay).

3. Likelihood of Returning to the Community—Patients were classified into two subgroups: those who were likely to return, and all others (this included the group who were in need of terminal care, as well as those who might return and those who were thought definitely to require permanent institutionalization).

4. ADL—The patients were classified into two groups: good ADL and poor ADL. The poor ADL group were dependent on at least five of the following six functions: bathing, dressing, toileting, transferring, continence and feeding. According to the work of Katz,[11] the most likely distinction between the two groups was continence, i.e., patients in the good ADL group were likely to be continent while patients in the poor ADL group were likely to be incontinent. We thus label the poor ADL group "incontinent."

5. Behavior—Patients were classified into two groups: those exhibiting frequent behavior problems of any type and those

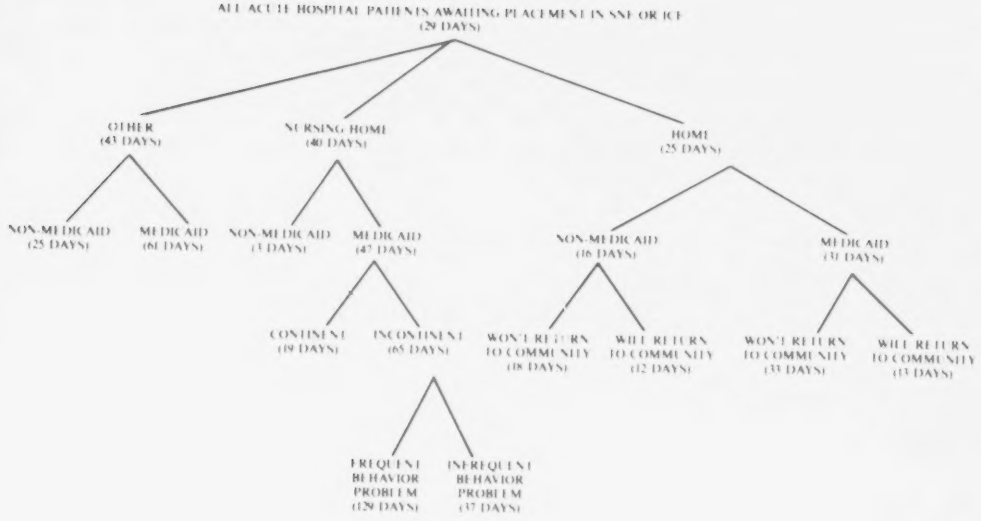


FIG. 1.

exhibiting no or infrequent behavior problems.

Using these five variables, we created ten distinct subgroups. The findings regarding the mean waiting time of these different subgroups in the exploratory sample are presented in Figure 1. The major findings are as follows:

1. For each referral source, there were striking differences between the mean waiting times of patients who would have different sources of payment in the nursing homes. The differences are most pronounced for the patients admitted to the hospital from nursing homes—Medicaid patients had to wait 47 days before being placed, while non-Medicaid patients were placed after an average of three days. Patients admitted to the hospital from home or from an "other" referral source also had a much longer delay in the hospital (roughly twice as long) if their stay in the nursing home was to be reimbursed by Medicaid.

2. For Medicaid patients admitted from nursing homes, there were two significant distinctions: patients who were incontinent took much longer to place than those who were continent; and among those who were incontinent, those with frequent behavior problems were much more difficult to place than those who did not have such problems.

3. For patients admitted from home, there was a common pattern for Medicaid and non-Medicaid patients: individuals who were likely to return to the community were placed much more rapidly than others.

After completing the analysis of the exploratory data set, we tested the reproducibility of these findings by examining the confirmatory data set. First, we studied the relatively refined distinctions between pairs of subgroups as shown in the lower "branches" of Figure 1 by determining whether the confirmatory data set yielded the same relative order for the mean waiting times of those pairs of subgroups in the lower branches. We found that in the case of one of these "contrasts," that between the two behavioral subgroups who were part of the nursing home–Medicaid–incontinent group, the mean waiting times reversed order in the confirmatory data set. For this reason, we collapsed these two subgroups into one.

After completing this analysis, we combined the exploratory and confirmatory data sets to test the significance of the observed differences in means. When differences were not significant, we merged the subgroups. In this way, the number of groups was reduced to five. The mean value and standard deviation of the waiting time of each of the five distinct subgroups are shown in Table 2.

Discussion of Analysis of Patient Characteristics

How Much Can Patient Characteristics Explain? Table 2 shows that it is possible to identify subgroups of patients with different profiles who will experience very different delay patterns. Moreover, the distinctions among these subgroups hold up when the data are subjected to a confirmatory analysis. This strongly suggests that the subgroups defined in Table 2 have a real meaning in that they relate to the preferences of nursing homes for different types of patients.

At the same time, there remains quite a bit of unexplained variation in the waiting times, as evidenced by the rather large standard deviations. Note that the standard deviation of the total waiting time of each subgroup in Table 2 is very close in magnitude to the mean value. This feature is characteristic of the geometric probability distribution.†

This suggests that although there are distinct subgroups of patients having very different expected waiting times, there is a great deal of variation in the amount of time any particular type of patient must wait. This random variation is not related to the patient characteristics we studied but is perhaps connected with other variables (hospital, area, nursing home) not included in this part of the analysis.

It is also possible to make quantitative estimates of the explanatory power of the patient characteristic variables. This was done by conducting an analysis of variance of differences in the total waiting time that can be explained by group membership. The ten-group model accounted for 27 per cent of the variance in the exploratory sample and for 11 per cent in the confirmatory sample. This dramatic drop attests to the wisdom of employing a separate confirmatory sample. The five-group model accounted for 12 per cent of the variance in the combined data sets. By conventional standards, accounting for 12 per cent of the variance is not a major accomplishment. However, the special nature of the placement process makes the 12 per cent figure respectable. In ordinary regression analysis or analysis of variance, the underlying theoretical model assumes that the error has a Gaussian distribution with constant variance throughout the range of the dependent variable. However, as we have already noted, waiting time is well described as a geometric random variable, for which the standard deviation nearly equals the mean. Thus, the percentage of variation accounted for depends strongly on the predicted mean waiting time, and there is a large, irreducible component of random error even for a model that classifies patients and estimates means perfectly. For instance, suppose that the five groups from Table 2 were in fact the basic subdivisions of the AD population and their relative proportions and mean waits were exactly as estimated. We show in Appendix B that, in this case, a perfect model would account for only about 15 per cent of the variance in waiting times. Thus, the 12 per cent figure represents about 80 per cent of the maximum achievable explanatory power.

Who Gets Placed and Who Must Wait? Additional insight into this question can be

† The geometric distribution describes the situation that occurs when dice are rolled. For a single die, the probability p of any particular number coming up is 1/6. The probability that you will throw the die n times without getting that number is given by (1-1/6) to the n^{th} power.

TABLE 2. Mean and Standard Deviation of the Total Waiting Time for the Five Distinct Patient Subgroups

Description of Subgroup	Mean	Standard Deviation	Number of Patients
A. Nursing home, non-Medicaid	3.3	3.1	15
B. Home, will return to the community, Medicaid and non-Medicaid	11.8	13.1	74
C. Other, non-Medicaid; other, Medicaid; nursing home, Medicaid, continent; home, non-Medicaid, unlikely to return to community	18.7	19.1	189
D. Home, Medicaid, unlikely to return to the community	38.4	45.9	213
E. Nursing Home, Medicaid, incontinent	50.1	58.7	84
Total	29.3	38.0	575

gained by further collapsing the five groups of patients into two larger groups, which we will call "easy" (for easy to place) and "hard" (for hard to place). This simplification costs little in predictive power, since the two-group model accounts for 10 per cent of the variance in the combined data sets, compared with 12 per cent for the five-group model. The 10 per cent figure represents about 70 per cent of the systematic variation that can be explained. The "hard" group comprises subgroups D and E in Table 2; these Medicaid beneficiaries have mean waiting times greater than the overall mean. Most of these "hard" patients (213 of 297) are individuals who were residing in the community prior to their hospitalization. They are regarded by the discharge planners as unlikely to return to the community, i.e., they are likely to become permanent nursing home patients. The remaining 84 patients who are "hard" are often called the "heavy care" nursing home patients. They are patients who were sent to the hospital from a nursing home; they are very dependent in ADL (incontinent) and paid for in the nursing home by Medicaid. After their hospitalization, the nursing homes do not want to take them back.

All other patients (identified as subgroups A, B and C in Table 2) are in the "easy" group, which includes all of the non-Medicaid and the remaining Medicaid patients. Note, though, that it is quite possible for an "easy" patient to endure a fairly long wait. The easy/hard distinction refers to average waiting time.

As we observed earlier (see section on estimation of ADs), the patients who are found in the queue present a biased picture of the waiting time problem that does not represent the situation of a typical hospital patient who becomes ready for discharge to a nursing home. It is very instructive to examine the composition (in terms of "easies" and "hards") of these two cross sections of patients. The numbers of patients and mean waiting times of these two cross sections are shown in Table 3. In this table, the "new" patients are those who became ready on the day of the survey, and the "old" patients are those who became ready at some earlier date and were already in the queue on the day of the survey.

The following observations are based on the figures presented in Table 3:

1. "New" patients are much more likely to be "easy" than "hard." The "easies"

TABLE 3. Mean Waiting Times of Easy and Hard Patients for New and Old Cross Sections

Patient Type	Cross Section		
	New	Old	All Patients
Easy			
No. and % of patients	69 (71%)	209 (44%)	278 (48%)
Mean waiting time (days)	6.2	19.2	16.0
Hard			
No. and % of patients	28 (29%)	269 (56%)	297 (52%)
Mean waiting time (days)	13.4	44.8	41.8
Total			
No. and % of patients	97 (100%)	478 (100%)	575 (100%)
Mean waiting time (days)	8.2	33.6	29.3

make up 71 per cent of the "new" group but only 44 per cent of the "old" group.

2. The waiting times of both "easy" and "hard" patients are more than three times greater in the "old" group than in the "new" group.

These observations provide a graphic demonstration of our claim that a snapshot survey exaggerates the length of time spent on AD status. People already in the queue are much more likely to be "hard" than "easy" and to have longer waiting times than the "new" patients who become ready on a given day.

It is also instructive to examine the differences between "easy" and "hard" to place patients against the backdrop of a snapshot photograph of patients who already reside in nursing homes. As we noted in the section on estimation of ADs, there is a significant difference in expected length of stay between a typical patient admitted to a nursing home and a typical resident of a nursing home. A comparison of the distribution of length of stay of these two groups is shown in Table 4. These data are from the 1977 National Nursing Home Survey.[10]

Table 4 shows that nearly two thirds of the discharged patients have a stay of less than six months; on the other hand, nearly three quarters of resident patients have been in the nursing home for more than six months. It is unlikely that any patient who has resided in a nursing home for more than six months is going to be discharged home.‡ Thus we can see that the bulk of *residents* of nursing homes are of the long-stay, custodial type. Most of these patients are paid for by the Medicaid program. Thus, it appears that the most typical nursing home *resident* (as distinct from admission)—the long-stay custodial care patient—is the type of patient nursing homes are least likely to want and is the type most likely to be backed up in the hospital.

Hospital Characteristics

We investigated the role of hospitals themselves in influencing ADs by correlating hospital characteristics with residual waiting times, i.e., with the differences between observed waiting times and waits expected under the two-group model. We used residuals rather than actual waiting

‡ In fact, according to data obtained from the National Nursing Home Survey, whereas 36 per cent of patients who are discharged from nursing homes after a stay of six months or less are discharged to their own or their family's home, less that 4 per cent of those patients discharged after more than six months return home.

TABLE 4. Comparison of Lengths of Stay of Residents and Discharged Nursing Home Patients, U.S., 1977

Length of Stay	Resident Patients	Discharged Patients
Less than 3 months	14.5%	53.6%
3–6 months	9.4%	12.0%
6–12 months	12.5%	10.0%
1–3 years	32.8%	14.0%
3–5 years	14.8%	6.1%
5 years or more	16.0%	4.1%
Total	100.0%	100.0%

times to provide some statistical adjustment for the varying mix of "easy" and "hard" patients across hospitals. The hospital characteristics we used were the size-adjusted occupancy rate and the number of patients found waiting by the survey. We chose these characteristics to test two theories about hospitals: that the AD problem will be worse in hospitals with low occupancy rates and that hospitals with many AD patients are "bad actors."

In the analysis using occupancy rate, we divided the 93 hospitals in the exploratory sample into six bed-size groupings. We then classified the hospitals into two groups, according to whether their occupancy rates were above or below the average for their bed-size category. This standardization was necessary because the occupancy rate has little normative significance unless related to bed size. Mean residual waits were three days longer in hospitals with below-average occupancy, but there was so much overlap between the two groups that the difference in mean residuals had no statistical significance ($Z = 0.8$, $p > 0.40$). If hospitals with low occupancy relative to their peers feel pressures to "hold on" to AD patients, the effect is not large enough to be reliably detected in our data set.

The second analysis focused on the relationship between mean residual waiting times and number of AD patients found in a given hospital on the survey day. The analysis revealed that the presence of large numbers of AD patients is not itself a sign of hospital complicity in the AD problem. A hospital with a large number of AD patients was not more likely to have a higher-than-average wait, after adjusting for the case mix, than a hospital with only a few AD patients (rank order correlation = 0.05, $P > 0.49$).

These findings do not necessarily mean that hospitals share none of the blame for the AD problem. The Massachusetts Department of Public Welfare[5] documented cases of halfhearted or inefficient placement efforts in a study of 14 Massachusetts hospitals. Nevertheless, it appears that neither low occupancy nor large numbers of AD patients are reliable indicators of hospital complicity.

HSA Characteristics

Besides patient and hospital characteristics, characteristics of the HSA in which the hospital is located might influence waiting times. Residual waiting times did vary by HSA, as shown in Table 5. Since the highest and lowest mean residual waits occurred in the HSAs with the highest and lowest combined occupancy rates for skilled and intermediate level facilities in 1975, it was natural to explore the relationship between HSA nursing home occupancy and residual waiting time. Our analysis used weighted regression methods to take account of differences among HSAs in the reliability of estimates of mean residual wait. We found a positive relationship between occupancy and residual waiting time:

$$\text{Residual Wait} = -462.3 + 4.922\ \text{Occupancy Rate},$$
$$R^2 = 0.65,\ F = 7.58,$$

But since this positive relationship depends so heavily on the two extreme

TABLE 5. Residual Wait and Nursing Home Occupancy by Health Service Area

HSA	Residual Wait			Nursing Home Occupancy
	Mean	SD	Number of Patients	
1	1.9452	37.8119	79	92.28
2	−4.8763	27.4472	81	93.66
3	2.5614	22.6969	41	93.54
4	−1.9277	38.3576	198	94.25
5	−13.0653	17.8162	61	91.65
6	11.4344	51.0310	115	94.68
Total	0.000	37.3902	575	

points, it should be viewed with some caution.

To get a more robust estimate of the relationship, we also computed a "jackknife" estimate,[12] which is based on calculations requiring the sequential deletion of each HSA from the data set. The jackknife estimate is calculated as follows:

$$\text{Residual Wait} = \underset{(1.88)}{-678.5 + 7.215\ \text{Occupancy rate}}$$

where the approximate t-statistic is shown in parentheses. This analysis, which is a more realistic depiction of a replication of our analysis, shows a positive relationship of marginal statistical significance.

It might be that some other characteristic of the HSAs is more fundamentally related to waiting time, with occupancy only a proxy. However, nursing home occupancy has an obvious a priori appeal and did relate positively to residual waiting time. Since nursing home occupancy across the HSA in which the hospital is located is a blunt indicator of nursing home bed availability, analyses using it will tend to underestimate the correlation with residual waiting time. Perhaps subsequent analyses will take more careful note of localized occupancy variations within HSAs and thereby discover an even stronger relationship.

Conclusions

This analysis suggests that the roots of the AD problem are not simple. In fact, our findings support two hypotheses: 1) as indicated by the sharp differences in the average waiting times of the different patient types, nursing homes are indeed admitting patients very selectively; and 2) the dependence of residual mean waiting time (i.e., mean waiting time after controlling for differences in patient-mix) on geographic area is an indication that the supply of beds has an effect on waiting times.

In addition, it is likely that the third explanation—that certain hospitals are dragging their feet—may also be correct. Although we have found that neither hospital occupancy rates nor the absolute number of patients waiting in the hospital correlates with the case-mix adjusted mean waiting time, the existence of large residual mean waiting times for a number of hospitals strongly suggests that sluggish hospital discharge planning is probably a factor in certain hospitals. What our results fail to establish is an association between foot-dragging and either low occupancy or large numbers of AD patients.

The results of this analysis, coupled with some experience with organization of long-term institutional services, can help

in developing strategies for reducing the undesirable effects of the AD problem.

We first note that our analysis leads to the rejection of some of the simpler proposals. For example, one such proposal is to expand the supply of beds. Although a significant expansion in bed supply might help to alleviate the problem, this would not be an efficient solution; the sharply different mean waiting times for different types of patients suggest that it would take a very large increase in bed supply to reduce the AD problem significantly. Whatever improvement in access would occur is likely to benefit those whose access is already best.

Similarly, proposals to penalize hospitals by significantly reducing their reimbursement for ADs will have limited impact, given the extreme reluctance of nursing homes to accept certain types of patients.

The strategy of paying nursing homes more to take care of certain types of patients (the so-called "heavy care" patient) may lead to a reduction in ADs, provided the change in reimbursement is sufficient to overcome providers' reluctance to accept certain types of patients. But note that differential reimbursement probably requires a simultaneous change in regulatory requirements regarding nursing home levels of care. (This is discussed in more detail in the "Discussion" section.) At the very least, a reimbursement-based initiative must take into account the fact that many of the "undesirables" who are backed up in the hospitals have similar profiles to many (if not the majority of) patients who are already residing in nursing homes.§

Discussion

The current regulatory system, which attempts to encourage nursing homes to group patients according to the level of nursing skill required in their treatment, ignores the very important differences among individuals in the type and total amount of care they need.[13] This has several important consequences, which exacerbate the AD problem. First, it must be noted that each level of care includes patients with a broad range of disabilities requiring widely disparate amounts of care. Nursing homes tend to select those patients who require a smaller total amount of care, all other things being equal. For this behavior, often called "creaming," providers are accused of being self-seeking and profit oriented. However, this behavior could be interpreted as a rational response to the regulatory/reimbursement system. Since most custodial long-term patients will require more care as time progresses, and since nursing homes are forced by reimbursement practices to work with fixed staffing patterns, the homes select less disabled new patients to balance an increasingly sick case load.

A second consequence of the "level of skill" definition is that it has fostered the mixing, in single nursing units, of patients having diverse and difficult problems—including dementia, frequent incontinence and various behavioral disorders—with patients whose primary limitations are physical. It is not surprising that many nursing homes are reluctant to accept patients who have special problems; it is, in fact, very difficult for both the

§ The state of Massachusetts is currently addressing the AD problem by providing augmented Medicaid reimbursement rates for nursing homes who accept AD patients. This policy does not take into account the similarity in profile of the typical AD patient and the typical nursing home resident. This policy appears to us to be ill-advised; it will create an incentive for nursing homes to refer more of their patients to acute hospitals, since, when these patients return, the nursing home may in some circumstances be paid at a higher rate. Furthermore, since the policy does not distinguish among patients by length of time on AD status, it encourages creation of ADs for easy-to-place patients, while offering no differential reward for nursing homes to admit hard-to-place patients.

staff and for the physically handicapped but mentally alert patient to deal with the problems caused by these difficult patients.

We believe, therefore, that current regulations that differentiate among patients according to the level of nursing skill needed in their care are not compatible with a rational organization of services in long-term care facilities. Modifications in level-of-care regulations and in reimbursement practices may facilitate a solution to the AD problem. The new regulations should allow facilities to group patients according to the total amount of care they need (instead of the level of skill needed) and to segregate patients having certain types of conditions into separate units (with separate reimbursement). A system such as this has been developed and successfully implemented in at least one large long-term care facility,[14,15] and the time appears right for a larger-scale experimentation with these concepts.

References

1. Shapiro E, Roos NP, Kavanaugh S. Long term patients in acute care beds: Is there a cure? The Gerontologist 1980;20:342.

2. U.S. Department of Health and Human Services, Office of the Inspector General. Secretarial report, Restricted patient admittance to nursing homes: an assessment of hospital back-up. August 1980.

3. Beatrice D. Briefing paper: Administratively necessary days. Internal report prepared for the Massachusetts Department of Public Welfare, December 1979.

4. Dumbaugh KA, Mackler RO. Report on patients in Massachusetts hospitals awaiting placement into long-term care. Massachusetts Hospital Association, October 1979.

5. Massachusetts Department of Public Welfare, Long Term Care Unit. The administrative day problem: an analysis of selected hospitals. July 1980.

6. Jordan H, Gruenberg LW. Trends in the problem of placing Massachusetts hospital patients in long-term care facilities. Presented at the 104th Annual Meeting of the American Public Health Association, October 1976.

7. Beattie RT, Jordan HS. A preliminary analysis of a survey of Massachusetts hospital patients who were ready for discharge and awaiting placement in other facilites on March 20, 1974. Massachusetts Department of Public Health Working Paper, April 1975.

8. Cox DR, Lewis PAW. The Statistical Analysis of Series of Events. New York: Wiley, 1966, 61.

9. Drake AW. Fundamentals of Applied Probability Theory. New York: McGraw-Hill, 1967;149.

10. National Center for Health Statistics. The national nursing home survey: 1977 summary for the United States. July 1979.

11. Katz S, et al. Studies of illness in the aged, the index of ADL: A standardized measure of biological and psychosocial function. JAMA 1963.

12. Mosteller F, Tukey JW. Data Analysis and Regression: A Second Course in Statistics. Reading, Mass.: Addison-Wesley, 1977;133.

13. Bishop CB, Plough AL, Willemain TR. Nursing home levels of care: Problems and alternatives. Health Care Financing Review 1980;2:33.

14. Besdine RW, Gruenberg LW, Kierman M, May MI. A cost-effective quality of life level of care system for the institutionalized elderly. Presented at the 107th Annual Meeting of the American Public Health Association, October 1980.

15. Besdine RW, Gruenberg LW, Frady L, May MI. A stratified care system promoting high quality, cost-effective institutional long-term care. Presented at the 108th Annual Meeting of the American Public Health Association, October 1980.

Appendix: Analysis of Mixtures of Geometric Distributions

We consider the magnitudes of the coefficient of determination (R^2) that will arise in regression analysis of waiting time distributions formed by mixtures of geometric random variables. Suppose that there are G groups, each containing N_i patients and having a geometric distribution of waiting time with mean M_i. Let the waiting time of the j^{th} member of the i^{th} group be X_{ij}, the sample mean waiting time for each group be $\overline{X}_i$, and the sample variance be S_i^2. Let the grand mean over all the groups be M_o and its sample value be $\overline{X}_o$.
By definition:

$$R^2 = 1 - \left[\sum_{i=1}^{G} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2 \Big/ \sum_{i=1}^{G} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_o)^2 \right]. \tag{1}$$

Using the identity

$$X_{ij} - \overline{X}_o = (X_{ij} - \overline{X}_i) + (\overline{X}_i - \overline{X}_o) \tag{2}$$

and the definition

$$S_i^2 = \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2 / (N_i - 1), \tag{3}$$

we can rewrite (1) as

$$R^2 = \left[1 + \sum_{i=1}^{G} (N_i - 1)\, S_i^2 \Big/ \sum_{i=1}^{G} N_i (\overline{X}_i - \overline{X}_o)^2 \right]^{-1}. \tag{4}$$

Finally, since the standard deviation and the mean are nearly equal for a geometric distribution and since for large groups the sample mean will be close to the true mean, we can use as a reasonable approximation

$$R^2 = \left[1 + \sum_{i=1}^{G} N_i M_i^2 \Big/ \sum_{i=1}^{G} N_i (M_i - M_o)^2 \right]^{-1}. \tag{5}$$

For instance, using the values of N_i and the estimates of M_i and M_o found in Table 2 of the text, we would expect regression analysis of a five-group mixture of geometric distributions to yield an R^2 value of approximately 0.15 at most.

MEDICAL CARE
February 1982, Vol. XX, No. 2

EPSDT Impact on Health Status: Estimates Based on Secondary Analysis of Administratively Generated Data

PATRICK H. IRWIN, PH.D. AND ROSEMARY CONROY-HUGHES, M.A.

EPSDT (early and periodic screening, diagnosis and treatment), a large-scale operational screening program that has generated a tremendous volume of data on the sociodemographic characteristics and health status of Medicaid-eligible children, seems to provide an ideal context within which to evaluate the effectiveness of preventive child care. Concern about health care expenditures generally, and the effectiveness of preventive child health services specifically, lead to the question of whether the impact on the health status of the children served can be measured without significantly adding to the cost of these services with primary data collection. We employ a quasi-experimental research design, using administratively generated data from an operational EPSDT program to estimate program impact on the prevalence of serious abnormalities among the children served. We found that, compared either with themselves across time or with a control group, a representative sample of 1,831 children had almost 30 per cent fewer abnormalities requiring care on rescreening.

Description of the Project

Background

RESEARCHERS in the field of child health have been unable to agree on the value of preventive services for children. However, in a recent federally sponsored critique of 38 controlled empirical studies, one reviewer argues that the credibility of the evidence on preventive care is significantly diminished by the poor methodological quality of much of the research. Although the studies are not totally inconclusive, the report concludes that "preventive child health care has, in general, been given only the weakest of tests."[1]

Evaluation of the long-term effectiveness of preventive care programs is a singularly difficult task, involving the costly and data-intensive process of measuring change over time. The increasing burden of health care expenditures suggests that such evaluation is essential to the development of cost-effective services. The costs of mass screening programs and other early-intervention measures can only be justified if those programs produce clear and lasting benefits. It is therefore imperative that evaluative evidence on the efficacy of preventive care be compiled and analyzed.

EPSDT (early and periodic screening, diagnosis and treatment), a large-scale operational screening program that has generated a tremendous volume of data on the

From the Philadelphia Health Management Corporation, Philadelphia, Pennsylvania.

This project was funded under Grant Number 18-P-97115, Health Care Financing Administration.

Address for reprints: Patrick H. Irwin, Ph.D., Philadelphia Health Management Corporation, 841 Chestnut St., Suite 1290.

0025-7079/82/0200/0216/$01.45 © J. B. Lippincott Co.

sociodemographic characteristics and health status of Medicaid-eligible children, seems to provide an ideal context within which to evaluate the effectiveness of preventive child health care. This study summarizes a methodology developed in southeastern Pennsylvania for estimating the impact of EPSDT; the results of this secondary data analysis, which employs a quasi-experimental research design, will undoubtedly be of interest to advocates and critics of preventive health services.

EPSDT in Southeastern Pennsylvania

EPSDT is a multiphase program of preventive health care for children of low-income families. The immediate goal of the program is to detect potentially debilitating health problems and, ultimately, to improve the health status and decrease the overall dependency of the target population. The screening procedure is *not* intended to be a complete diagnostic examination; rather, the screen serves to identify potential health problems and to distinguish those conditions that warrant immediate care or referral from those requiring no more than a simple follow-up. In Pennsylvania, all Medicaid-eligible children are entitled to be screened every three months for the first 18 months of life and every other year thereafter, through the age of 20.

The EPSDT program in the five-county Philadelphia area is administered by a private organization under contract to the State Department of Public Welfare and consists of five phases:

I. Outreach or case finding, to identify and contact eligibles, to inform them of the program and to assist them in gaining access to services;

II. Administration and coordination of screening, to establish and direct a network of qualified screening units to serve the area's target population;

III. The actual testing or screening by certified providers of medical care;

IV. Compilation and reporting of screening results; and

V. Follow-through and diagnosis, to secure provision of indicated treatment services when necessary.

Over the past seven years of its operation, a management information system (MIS) has been developed, and large quantities of managerial data have been produced. By 1980, over 700,000 screens had been perfomed in this system and the medical records stored in a large computer file.

However, these administratively generated data are not ideally suited for research and evaluation, and they suffer from serious problems of data quality and completeness. The eligibility status of the target population changes rapidly, and there is an unavoidable time lag before a change in eligibility is posted to the data system; the program's MIS therefore does not provide an accurate estimate of the current number of eligibles at any point in time. Although screening data tend to be more complete than eligibility data, even this most reliable component of the MIS contains numerous instances of misidentification, duplicate records and anomalous data values. Finally since a relatively small proportion of children is screened more than once, longitudinal data are available for a very limited number of children.

Objectives of the Present Study

This study was undertaken as part of a larger effort to examine the process, effectiveness and costs of the EPSDT program in southeastern Pennsylvania, and it represents an attempt to determine the extent to which EPSDT has had an ameliorative impact on the health status of children receiving periodic screens (i.e., the program's effectiveness). As a demonstration, the project attempted to show to what extent, and how, the administratively generated data base could be exploited to estimate that effectiveness.

The project's initial attempts to measure the impact of the program on health status using only available data seemed to give ambiguous results. In terms of change in risk status between screen and rescreen (i.e., in the proportion of children having a serious problem on first screen compared with their second screen), no overall change is apparent in a sample of children who received two screens approximately two years apart (See Table 1).

The McNemar test results in a nonsignificant chi-square of 400/783 = 0.51 (df = 1, p = 0.48) and indicates that no *net* change in risk status is evident in this sample. Similarly, there is no evidence of significant changes in risk status whether or not the child returns to the same provider (nearly 58 per cent did) for rescreen. (See Table 2; chi-square is 0.037 and 1.786, respectively, with $p > 0.10$ for both.)

Tables 1 and 2 illustrate the impact of EPSDT by focusing on turnover—the clients who change risk status, from having a problem to not having one, and vice versa. In Table 1, 783 (402 + 381) children (42.7 per cent of the sample) changed status, even though the observed *net effect* over time is only approximately a 2 per cent decrease in those with problems. (Again, since it is not significant, the change in the population could be zero or even in the opposite direction for the entire population.) We may therefore conclude that there is no evident impact on health status.

From another perspective, in terms of the persistence of specific problems, the curative impact can be quite clearly estimated with the same sample. Of the 992 children with problems that require care at time 1, 40.5 per cent are problem-free at time 2. This improvement is probably *under*estimated, since we do not know from Table 1 whether the 590 children with problems at both times still suffered from the same abnormality identified at time 1 (nor the same number of abnormalities).

Deriving a reliable estimate of risk reduction, to determine if the same abnormality were present at both times, requires considerable data manipulation and assumptions about their meaning. Short of a medical record audit, we are able to determine whether the problems identified on rescreening were in the same general diagnostic area as those identified at the initial screen; if the subsequent problems are not in the same area, then they can be reasonably assumed not to be the same problem. Table 3 summarizes the results.

For the sake of presentation, we distinguish between risk reduction and problem resolution. A child is classified as "high risk" if there is at least one problem requiring care identified on screening. It can be argued, however, that risk is reduced if at least one of those problems is resolved. Surely some children classified as high risk at both times will have enjoyed reduced

TABLE 1. Change in Risk Status from Initial Screen to Rescreen

		Time 2	
		No Problem (0)	Problem (1)
Time 1	(0)	450	381
	(1)	402	590

TABLE 2. Change in Risk Status from Initial Screen to Rescreen for Same Provider and Different Providers

		Same Provider	
		Rescreen	
		No Problem (0)	Problem (1)
Initial screen	(0)	255	219
	(1)	214	366
		Different Provider	
		Rescreen	
		No Problem (0)	Problem (1)
Initial screen	(0)	203	162
	(1)	188	224

risk before rescreening but will have contracted another or a similar problem. In Table 3, a problem that "persists" (32.3 per cent of the total) may in fact not be the same abnormality but rather another problem in the same diagnostic area—another source of impact underestimation.

Undoubtedly, some problems requiring care will persist because the conditions are chronic. The remaining conditions that go untreated may be few indeed, which leads to a tentative conclusion that in terms of potential *curative* impact, the results are consistent with a significant improvement in health status among the EPSDT children in this sample.

While these conclusions from different approaches to the same data are not completely contradictory, they are ambiguous about the impact of EPSDT on health status. We attempt to resolve this ambiguity by developing a more complex research design, which incorporates longitudinal as well as cross-sectional comparisons, yet does not necessitate primary data collection.

Research Design

Conceptual Definition of Impact

Despite the limitations of the data system, it should be possible to determine, using only data generated by the program itself, the impact of EPSDT on the health status of the children it has served. Numerous confounding factors make this determination complex; nonetheless, it is important to know to what extent the impact can be isolated, without resorting to primary data collection or to experimentation with human subjects. "Impact" in this instance refers to a change in health status attributable to a given process (viz. EPSDT) and must be distinguished from those program performance measures that evaluate cross-sectional differences between groups rather than longitudinal change.

Changes in health status can be determined only for a specified individual or group passing through a given process. Differences, on the other hand, can be observed between or among any units of comparison. It is important to maintain the distinction between longitudinal changes and cross-sectional differences in this enterprise, as elsewhere.[2] The central methodological thesis of this study is that to analyze the impact of EPSDT on the health status of clients without collecting new data, it is necessary both to compare initial screen results with rescreen results *for the same children* and to compare children who have participated in the program with those who have not.

TABLE 3. Health Status at Time 2 of Children Who at Time 1 Had Had Problems Requiring Care (By Diagnostic Area)

Diagnostic Area	Problem Resolved		Problem Persisted		Total Problems
	n	%	n	%	
Dental	308	70.3	130	29.7	438
Vision	189	53.5	164	46.5	353
Hearing/speech	52	80.0	13	20.0	65
Development/behavior	41	85.4	7	14.6	48
Nutrition/growth	24	85.7	4	14.3	28
Medical	332	69.7	144	30.3	476
Total Problems	946	67.2	462	32.3	1408*

* Sums to more than the total number of problem screens (992) because 19.3 per cent of the children had multiple problems.

Central to this inquiry is a sample of children who received two screens approximately two years apart. Even if a simple comparison of health status (to be operationally defined below) at both times were to reveal improvements for this group of children, the differences could be due to something other than the benefits of the EPSDT assessment and referral. These rival explanations for observed differences in health status challenge our ability to attribute observed effects to the EPSDT program, but the challenge is a serious one *only if* the alternative explanations are plausible. The plausibility of competing explanations will be tested by comparing the two-screen sample with itself across time and with a control group of children just entering the program. To the extent that these comparisons tend to disconfirm any rival explanations, inferences about program impact are valid.

Comparative Design

Of the many assumptions that are necessary in this analysis, the most heroic is that impact—any observed change in health status—is due only to EPSDT. This admission is notably unsurprising. Outside (or inside!) a laboratory, unequivocal attribution of cause and effect is not a simple task. Still, if EPSDT as a health services program has been ameliorative, then we can reasonably expect some evidence of the effect. It is also reasonable that the possible sources of bias in available data be considered explicitly.

In any research endeavor, success in attributing observed (and hypothesized) change to some experimental treatment (e.g., the impact of a program) is primarily a question of the internal validity of the study. (The ensuing treatment of validity follows that of Campbell and Stanley.[3]) Did the program produce the change, or is there some alternative explanation (referred to as a "threat to the internal validity")?

We propose to examine explicitly the threats to internal validity that challenge this impact analysis. We begin, therefore, by repeating the eight primary threats to internal validity identified by Campbell and Stanley.

1. History—a specific event occurring between first and second measurement besides the experimental variable;
2. Maturation—changes occurring in the respondents as a function of time;
3. Testing—the effects of taking a test upon the scores of a second testing;
4. Instrumentation—changes in the observers or scorers producing changes in the obtained measurements;
5. Statistical regression to the mean—operating where groups have been selected on the basis of their extreme scores;
6. Selection—biases resulting from differential selection of respondents for experimental and comparison groups;
7. Experimental mortality—differential loss of respondents for the comparison groups;
8. Selection-maturation—and similar interaction effects.

Each of these sources of bias seems more or less plausible in the context of EPSDT impact analysis, with one significant exception. Testing cannot be considered as a threat to the internal validity of a research design in which the measurement instrument (the EPSDT screen) *is* the experimental treatment.

Random assignment of subjects to experimental and control conditions may be the optimal design with which to assess the impact of social programs, but this decisive, single-experiment approach is not feasible in most field settings.[4] Without the use of randomization, quasi-experimental research designs can only partially resolve the issues of threats to internal validity. Fortunately, in contrast to the conduct of research in an experimental environment, research conducted in situ offers a high potential for findings that can be readily ap-

plied or generalized to other operating environments (i.e., it offers high external validity).

The potential for application across contexts is crucial in situations in which inquiry will always be subject to challenges to internal validity. If the same impact is found in several contexts or in programs that are each subject to different uncontrolled variable effects, then the evidence for the single explanation of program impact grows "just because the plausible rival hypothesis is different from study to study."[3] This inferential feature is deliberately introduced in the following design. The design also attempts to provide not only a way around the ethical impasse to human experimentation by denial of indicated treatment, but also to demonstrate a cost-effective and unobstrusive mode of empirical research in preventive child health services.

Comparison Groups

Overall, the research design is comparative, with observations made both cross-sectionally (to measure differences) and longitudinally (to estimate changes). The groups to be compared are a two-screen sample (the experimental group), and two other samples of initial screens approximately two years apart.

Using the Campbell and Stanley diagrammatic convention, the research model may be formulated as follows:

Two-Screen Experimental Group: O1 X O2
One-Screen Control Group (1977): O3
One-Screen Control Group (1979): O4

The "O" stands for an observation at a given time, "X" for experimental treatment. Technically, in this case, the observation is the EPSDT health screen, and the treatment is the referral. The EPSDT assessment and referral are separable only conceptually, however, since our concern is the outcome of exposure to an EPSDT screen. The comparison is between those who have had this exposure previously and those who have not.

Sample of Rescreened Children. For a child to be included in the two-screen sample for this analysis, several criteria have to be met: 1) the child's first screen occurs between July 1, 1977, and December 31, 1977; 2) the child is at least 19 months of age at the time of the initial screen; and 3) the time interval from first to second screen is at least 18 (and up to 32) months. (After the restrictions to the sampling frame were imposed, only 1,831 children qualified, so all were included in the two-screen group. Thus, these children do not constitute a simple random sample, although this virtual enumeration is no less representative than a random sample.)

Criterion 1 insures first that these children all represent first-to-second rescreens rather than second-to-third, etc. Using a more homogeneous sample should improve the precision of our estimates of risk reduction. The beginning of the six-month time frame selected for the first screen corresponds to an internal change in the data collection system (in July 1977) that changed certain data items and codes.

Criterion 2 insures that we have not selected children who are 18 months of age or less and consequently on a special infant rescreening schedule. These children receive a somewhat different series of tests and are permitted to receive six screens from birth through the age of 18 months. Our interest in this exercise is focused on children who are eligible for the two-year schedule.

Criterion 3 insures that the two-year schedule has been followed at least approximately. Again, this should help to provide a more homogeneous sample.

Control Groups. The first control group (the O3 sample of the preceding diagram) consists of a sample of all children receiving their first screen at about the

same time as the two-screen sample in August 1977, and who were also at least 19 months old (n= 2889). This group overlaps with children in the 01–02 sample—those of the two-screen group who received their initial screen in August 1977. Some of the comparisons to be made assume the independence of samples and require the exclusion of this overlap; we shall refer to the restricted 03 sample as 03′, n = 2,578. (Again, for both control groups, limitations to the sampling frame, cost and convenience led us to include all who qualified in these nevertheless representative groups.)

The second control group (04) consists of children in the same age range who receive their first screen at approximately the same time that the two-screen sample was being rescreened, between January and April 1979. An effort was made to further restrict this group to those children from families who were new to the program as of January 1979, to insure sample independence (n= 1,182).

Operational Definition of Impact

A complete EPSDT screen in Pennsylvania includes as many as 39 individual Test Areas (TAs), organized her into six mutually exclusive summary Diagnostic Areas (DAs): medical, vision, hearing, dental, laboratory and behavior. (This study does not explicitly consider assessment of the self-reported immunization status, which is also part of the screen.) Some of these TAs may not be assessed in every screen because they are discretionary or are not appropriate for a given child.

The response categories are not identical for all TAs, but the alternatives provided for the medical evaluations are typical:

Code	*Response*
1	Normal
2	Abnormal, treatment required
3	Abnormal, nontreatable
4	Not done

For the purpose of this analysis, only code 2 (abnormal treatment required) is taken as evidence of a "serious" problem. Whether the child is referred or not is apparently more indicative of a provider's willingness, specialization or experience in treating the condition rather than a medical judgment or assessment of the child's health status. Codes 3 (abnormal, nontreatable) and 4 (not assessed) are counted as missing values.

For the longitudinal and cross-sectional comparisons required in the present study, we operationally define Health Status (HS) in terms of the total number of abnormalities that require treatment. Again, abnormalities that are identified as not requiring care or as not treatable (a relatively rare occurrence) are excluded. To standardize HS for comparisons among different groups of children, the total number of Abnormal Test Areas (ATAs) found is divided by the total number of test areas that are assessed for a group of children. A low score on this index therefore indicates a healthy individual or group; a higher score indicates a relatively higher incidence of illness or injury.

The measure of HS is calculated according to the following formula and can be expressed either as a proportion or a percentage:

$$HS = \frac{ATA}{TTA - NA}$$

Where, for a specified group of children, HS = health status index; ATA = total number of *abnormal* test areas where treatment is required; TTA = TAs × S, the number of test areas in which a treatable abnormality can be found (39) times the number of children screened; and NA = total number of test areas not assessed, an adjustment for a given group of children to eliminate TAs not assessed.

With a slight modification of this formula, we can also confine the focus of the measure to a particular DA; i.e., any

specified selection of TAs. For example, if we wish to consider only conditions within the dental DA, then the potential number of TAs is limited to three possible tests: caries, oral infections and other dental problems. When a particular subset of TAs (< 39) is selected for calculating the index, we shall refer to it as an Abnormality Rate (AR).

The HS index can be interpreted as the probability that the EPSDT screen will uncover any condition requiring treatment. Similarly, an AR for a specified DA can be interpreted as the probability that a specific kind of problem will be identified.

Adjustment Estimation Procedure

Before controlling for the relavant threats to internal validity, there seems to be virtually no change in HS for the longitudinal sample between screen and rescreen, and no difference in HS for the cross-sectional samples between rescreen and control group. For the unadjusted longitudinal sample, 2.61 per cent of all TAs assessed on initial screening in 1977 were abnormal, with a condition requiring treatment; that percentage appears to have *changed* by 1979 to 2.69 per cent on rescreening. This latter health status level is also compared with 2.58 per cent for the control group of children who received their initial screen in 1979.

To conclude at this point that there is evidently no impact would be premature. Particularly in light of the need to control for rival alternative explanations, we maintain that these unadjusted comparisons are invalid: the groups are not yet comparable.

In the sections that follow, we describe in some detail the means of estimating the needed adjustments and then compare the groups again after adjustment; actual adjustments are calculated in the "Findings' section. The procedures for estimating and adjusting may seem complex, but the conclusions about EPSDT impact on health status that we can make at that point will differ significantly from preceding conclusions based on unadjusted HS. While not all sources of incomparability of groups can be controlled in each comparison made, by integrating two modes of comparative analysis—longitudinal comparison of change with cross-sectional comparison of difference—we can gain a new perspective that is essential for evaluating impact based on this administrative data base.

After estimating the magnitudes and directions of each of the adjustments needed, the adjustment procedure is 1) to ascertain the per cent of change or differences due to factors other than the program—the threats to internal validity—and then 2) to add or subtract this percentage to/from the per cent of difference or change in the unadjusted HS rates. For example, if there has been a 10 per cent increase in abnormalities due to instrument changes, we would subtract this figure from the per cent change in HS for the longitudinal comparison between screen and rescreen. The effects of this procedure will be tabulated both for the overall HS index and the selected DAs.

Instrument and History Effects. To gauge changes in health status over time, the first requisite is that the assessment (instrument) remain constant. A less rigorous procedure at rescreen than at initial screen, for example, would probably result in fewer abnormalities being found and could mislead us into accepting an apparently improved health status that is actually due to instrument change. When there are events extraneous to the EPSDT program, which clients commonly experience between screenings, these events could alter health status (history). The introduction between screens of a new child health initiative for Medicaid-eligible children (e.g., an intensified nutrition or immunization program within the public schools) could mislead us into attributing improvements in health status to EPSDT rather than to history.

Instrument or history effects on health status assessment are analogous to the effect of inflation on income: health status may appear to improve across time only because its measurement is less rigorous. As with inflation, it is possible to control statistically for differences due to changes in instrument or history, but only if they can be identified and isolated from changes attributable to the program.

A test of the instrument/history hypothesis can be made by comparing the two control samples of initial screen results—in terms of the diagrammatic convention, the 03–04 comparison. Since the 03 and 04 control groups are independent and representative samples of chidren being screened at two times, differences between them can be attributed to instrument or history effects. Those observed differences that are also confirmed by program history as resulting from protocol or programmatic changes can be used to estimate the magnitude of the effect and to adjust apparent changes in the rescreened (i.e., experimental) group.

Of the 39 individual comparisons included in the HS index, nine revealed differences in ARs that are significant ($p < 0.05$). In general, we found higher ARs in the later (04) sample; this finding suggests that the screening protocol may have become more rigorous over time.

EPSDT program administrators have offered plausible explanations for virtually all of the observed differences. Administrative personnel attribute the differences in ARs to changes in EPSDT procedures or to changes in the Medicaid service delivery system that have occurred in this area in the two years between the screens. Thus, examination of independent groups from time 1 and time 2 suggests that factors relating to instrument and history may indeed threaten the internal validity of our longitudinal comparisons. (See Table 4.)

Table 4 presents the HS index and selected ARs for the 1977 and 1979 cross-sectional samples (03 and 04, respectively). The overall HS in the 1979 sample is approximately 21 per cent higher than the corresponding rate in the 1977 sample. This estimate of instrument effect will be used below to adjust the per cent change in HS between 01 and 02. (ARs are adjusted similarly according to their respective 03–04 differences.)

[Because the age distributions of the 03 and 04 samples are not the same, any AR differences between the groups may be attributable to age differences rather than to instrument changes. To age-adjust at this stage would introduce a new confound, but to avoid the possibility of confusing age differences with true instrument/history effects, we compared the ARs in two other

TABLE 4. Selected First Screen Abnormality Rates (ARs) in 1977 (03) and 1979 (04)

		03		04		
DA (# TAs included)		Total Tests Done	AR%	AR%	Total Tests Done	p-value
HS	(39)	83,608	2.13	2.58	34,407	0.00
Medical	(13)	37,542	1.30	1.96	15,375	0.00
Vision	(2)	4,508	9.05	9.96	1,969	0.26
Hearing	(2)	4,502	1.38	1.99	1,956	0.09
Dental	(3)	8,667	7.03	6.26	3,549	0.12
Laboratory	(10)	12,635	0.87	1.25	5,436	0.02
Behavior	(9)	15,754	0.67	1.00	6,122	0.02

Note: Indicated p-values are the two-tailed significance levels resulting from a difference-of-proportions test.

cross-sectional groups that, while not carefully selected samples of first screens, are already age-comparable. Using administrative statistical reports for July through December 1977 (covering 36,415 screens) and for October 1979 (4,424 screens), we found differences similar to each of those reported here.]

The estimated difference may reflect increasingly rigorous quality control and standardization of tests (e.g., "not assessed" vision TAs decreased from about 22 per cent in 1977 to 17 per cent in 1979). In addition to the HS adjustment, specific TAs vary in the amount and significance of difference. Therefore, any screen–rescreen change in HS or ARs (except Dental) toward improvement will be underestimated and therefore should be increased by the same amount. (Dental AR change will be overestimated because the probability of finding a dental abnormality decreased, although not significantly.)

Selection and Statistical Regression Effects. In gauging impact by cross-sectional comparison of rescreen results with the initial screen results of a control group, we realize that observed differences in health status could also be due to differential selection or statistical regression. If those who received two screenings represent another, perhaps healthier, population, selection bias or statistical regression to the mean may explain any cross-sectional differences.

A test of this rival hypothesis that is possible with the available data involves comparison of the longitudinal sample at initial screen with others being screened initially at that time; i.e., the 03′ control group. If those children who eventually received rescreening are significantly different (e.g., healthier) from others at the outset, a selection bias is indicated. The 01–03′ difference can be used to estimate the magnitude of the bias and to adjust the cross-sectional comparison of the rescreen results (02) and control group results (04). If the rescreened group enjoys the same health status as others being initially screened at that time, we know at least that the rescreened children did not have a head start.

Unfortunately, since there could be no random assignment from all children who were screened once to the group of children who were rescreened, rescreening may itself be taken as evidence that the children are generally different (i.e., extreme) in terms of health, specifically in tendency to improve, a selection–maturation interaction.[8]

Tendency to improve cannot be isolated easily using a longitudinal control group, observed both at time 1 and time 2 but not treated, because the health assessment (or experimental observation) is the "treatment." Still, by extending the comparison at time 1 of the initial screen children (01–03′) to nonhealth factors (e.g., demographic comparisons of available data), we might still support the claims that the rescreen group is equivalent to the control group at initial observation.

Receptiveness to intervention is a characteristic not necessarily discernable from initial HS; but some exogenous sociodemographic factors might present children with treatable abnormalities an environment that potentiates the effects of the EPSDT intervention. In other words, the chances that any indicated treatment regimen will be followed may be higher in certain households and among certain classes of children than in others.

Table 5 displays the summary DAs for 01 and 03′, including general HS. Generally, the two-screen sample has a significantly ($p < 0.0001$) higher abnormality rate at first screen (26.1 per cent higher) than does the control group. The medical, vision and dental DAs are the chief contributors to this difference.

Rather than enjoying a head start as compared with others receiving their initial screening at the same time, the lon-

TABLE 5. Selected First-Screen Abnormality Rates (ARs) in Two 1977 Samples

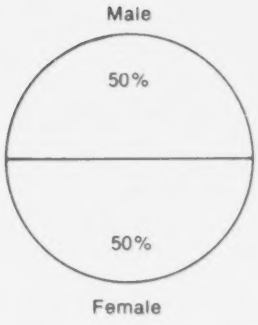
DA (# TAs included)		01 Total Tests Done	01 AR%	03′ AR%	03′ Total Tests Done	p-value
HS	(39)	53,446	2.61	2.07	97,929	0.00
Medical	(13)	23,801	1.58	1.27	33,500	0.00
Vision	(2)	2,975	11.97	9.03	4,010	0.00
Hearing	(2)	2,975	1.68	1.30	3,997	0.20
Dental	(3)	5,493	8.47	6.66	7,734	0.00
Laboratory	(10)	8,392	0.94	0.84	11,279	0.46
Behavior	(9)	9,807	0.71	0.68	14,009	0.78

Note: 03′ consists of all August 1977 initial screens *excluding* those belonging to 01 sample. Indicated p-values are the two-tailed significance level resulting from a difference-in-proportions test.

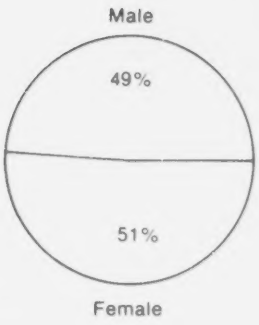
gitudinal sample appears to have had proportionately more serious problems. There are several possible reasons for this selection effect. One is that the identification of risk increases the probability that a child will be rescreened (i.e., the presence of a treatable abnormality on initial screen co-occurs with an increased incidence of rescreening of almost 29 per cent). While almost 60 per cent of the 03′ group were known to be eligible for EPSDT through May 1979, there is no record that any of them was rescreened by March 1980. Based on the entire 03 sample of 2,889 children who were screened for the first time in August 1977, 1,428 had problems, and 1,461 did not. Of those with problems, 173 or 12.1 per cent were rescreened, compared with 138 or 9.4 per cent of those without problems. The association between whether or not a child has a problem and whether or not the child will be rescreened is significant (chi-square = 5.32, df = 1, $p < 0.05$).

Another possible reason for the higher 01 ARs is seasonality, or the fact that the longitudinal sample is drawn from a six-month sampling frame as compared with a one-month frame (August) for the comparison group. Other possible reasons for the 01–03′ difference may lie in the differences in their age distributions (chi-square = 82.9, df = 17, $p < 0.05$) and in differences in the distributions of the providers who screened them (chi-square = 323.05, df = 74, $p < 0.05$). (Both chi-square statistics for age and provider differences are calculated after excluding categories with expected frequencies of less than 5.)

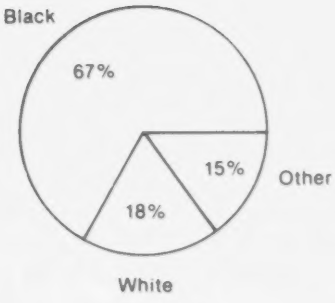
In order to test the possibility of a selection–maturation interaction or tendency to improve, the two groups (01 and 03′) were also compared on a series of demographic variables. In terms of sex, race and whether the child is from Philadelphia or the suburban counties, there are no significant differences. In terms of payment category ("categorically needy" or "medically needy" versus "general assistance"), household size and age distributions, the differences are statistically significant ($p < 0.05$) but otherwise very small. (See Fig. 1A and 1B.) [Other socioeconomic comparisons can be made of measures derived from 1970 Census tract-level data for most of the children in the samples. Given the nature of the data, these comparisons are extemely tentative, but they may provide a useful sense of environmental factors. The average (mean) median family income differs significantly between groups ($6,337 for 01 and $6,606 for 03′, $p < 0.05$), while



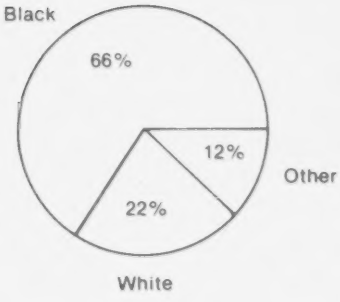
Sex-01



Sex-03'



Race-01



Race-03'

County-01

County-03'

FIG. 1A. Demographic comparisons at first screen between those who received a rescreen (01) and those who did not (03').

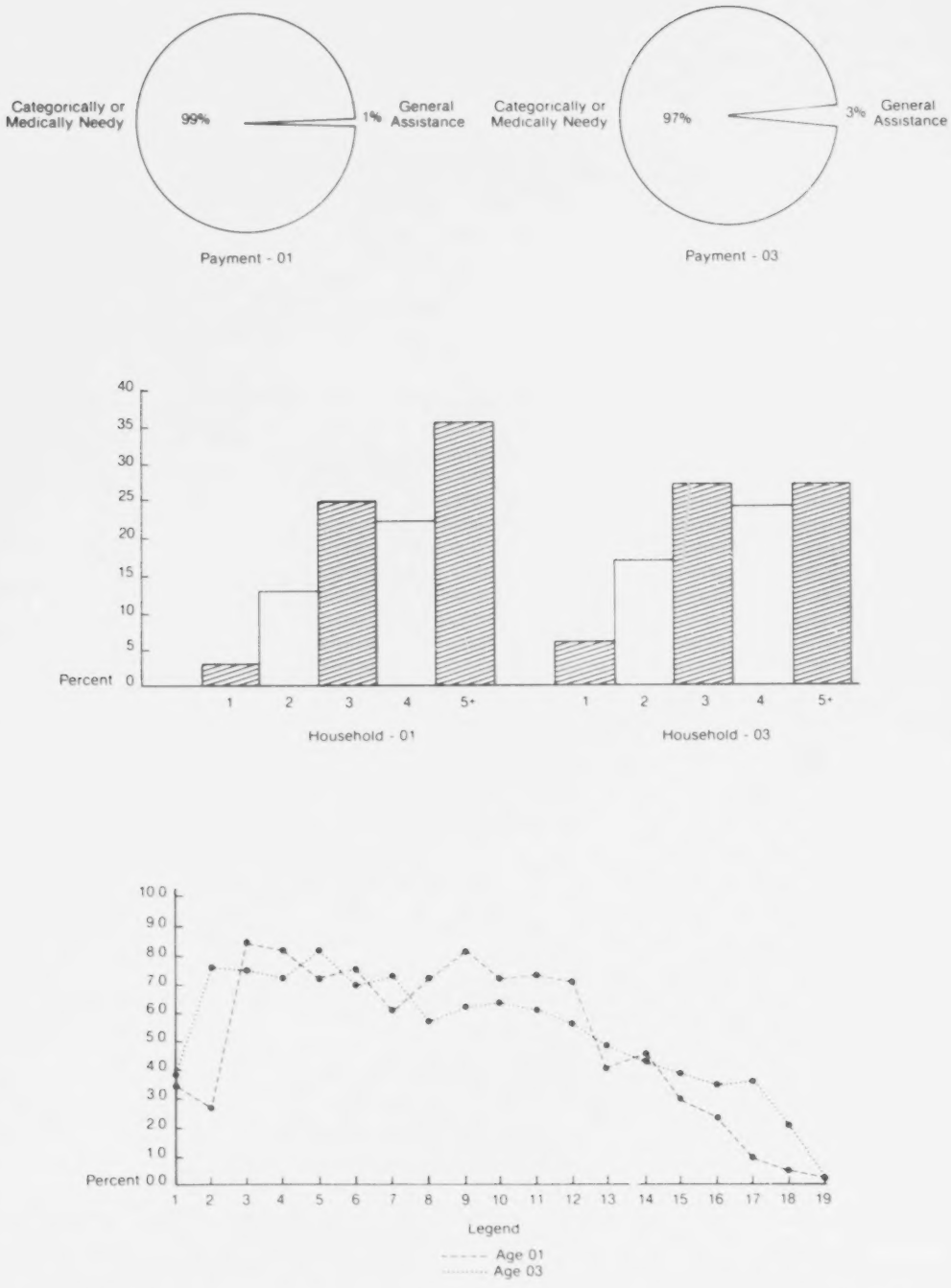


FIG. 1B. Demographic comparisons at first screen between those who received a rescreen (01) and those who did not (03').

the mean per cent living in crowded housing does not differ significantly (18.1 per cent and 19.8 per cent, respectively).]

Generally, the demographic differences described are interesting and suggestive, but, in terms of practical significance in support of the selection–maturation interaction hypothesis or as a source for an adjustment to the longitudinal comparison, they are too tentative. Moreover, we note that the coeval cross-sectional comparison does not suffer from this interaction threat to validity, as does the longitudinal comparison. If the results of both comparisons are the same, after other threats are controlled, tendency to improve would lack support as a rival hypothesis.

Maturation Effects. Another of the plausible alternative explanations for observed longitudinal changes is maturation. Campbell and Stanley[3] define maturation as "processes within the respondents operating as a function of time." As such, maturation is a form of change and can affect only the longitudinal (01–02) comparison. Inferences drawn from cross-sectional comparisons (02–04), on the other hand, may be invalidated by age differences between the experimental and control groups. An age-adjustment procedure must therefore be used to control for both maturation and age differences.

The mechanics of age adjusting involve a procedure analogous to deriving standardized specific rates (e.g., age-specific birth or death rate). The age standardization recalculates HS (or specific ARs) as if the two groups were identically composed on age. The formula is simply a weighted mean that expresses the age-specific HS of the two first-screen groups (01 and 04) weighted by the age distribution of the rescreened children (02).[5] If there are no age differences between any two comparison groups, then standardization will not change the HS, although the converse is not necessarily true.

Since no child in the rescreened group is under three years of age when they are rescreened, and none over 18 in the first-screen groups, it is important to note that standardization will limit the comparisons to only those children 3 through 18 years of age. (See Fig. 2A-C.)

Experimental Mortality. The validity of comparing rescreened children's health status with that of children just receiving their initial screens (02–04) is also threatened by the fact that, if the program's future emulates its past, most of the children in the comparison group (04) will not be rescreened. Since there are differences at first screen between those children who are eventually rescreened and those who are not (the selection effect), we may therefore suspect that, had all first-screen children received a second health assessment, our estimation of EPSDT impact would be different. The differential loss of children from the time series implies that the observed cross-sectional 02–04 differences may be due to "experimental mortality" rather than to EPSDT.

Longitudinal comparisons (01–02), on the other hand, do not suffer from the effects of differential selection, since the same group is compared intact across time. Therefore, although this research design requires both cross-sectional and longitudinal comparisons, only the inferences from cross-sectional data are subject to the challenge of experimental mortality.

The plausibility of experimental mortality as an alternate explanation for observed cross-sectional differences might be tested by locating and screening those children who were not yet rescreened. If this new estimate of health status for rescreened children does not differ from that of the previous rescreened group, the hypothesis of experimental mortality could not be supported. Even if the test were feasible, however, the resources required to locate and rescreen these children would be prohibitive.

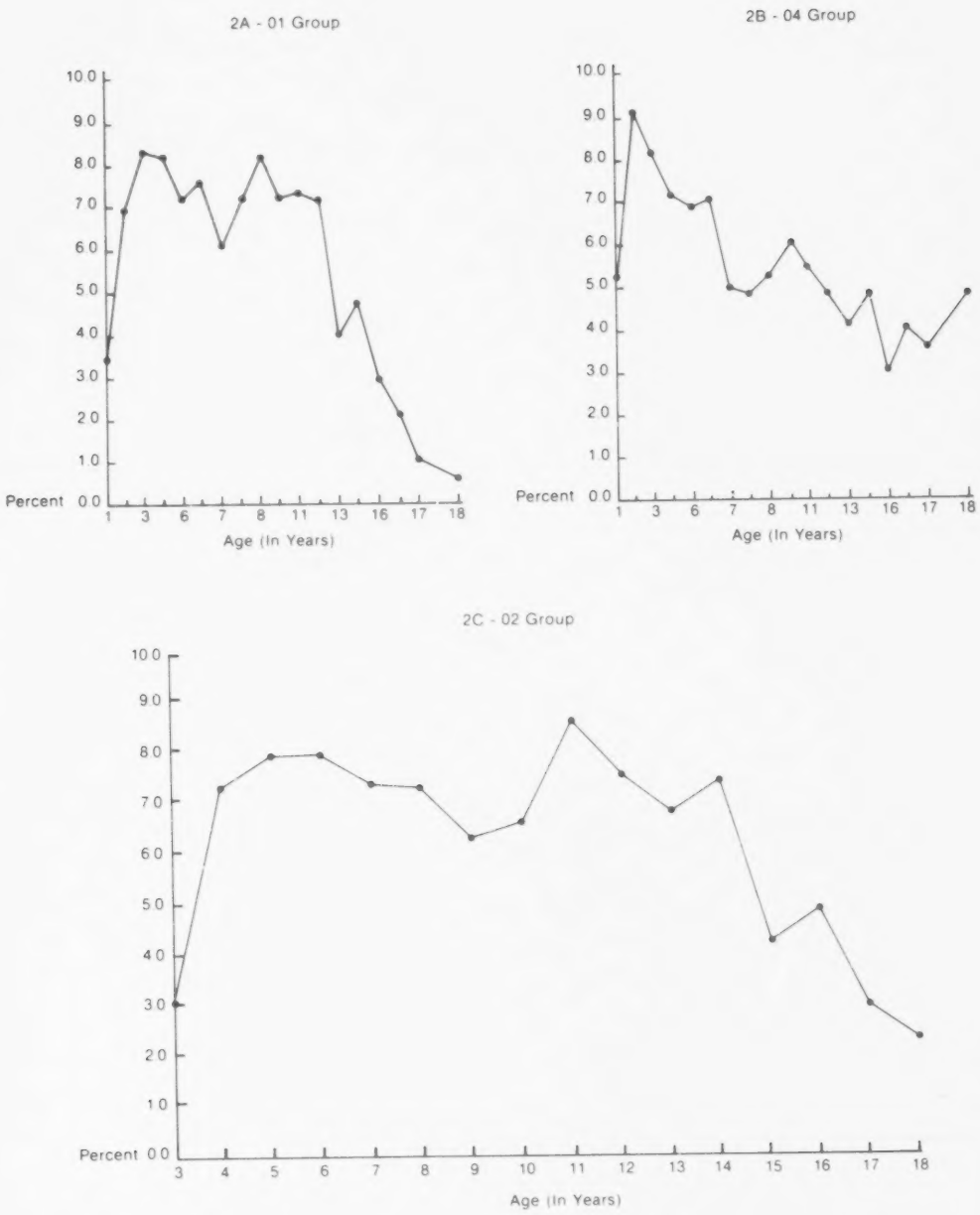


FIG. 2. Age distributions among the comparison groups. 2A, 01 group; 2B, 04 group; 2C, 02 group.

Given the practical difficulties of eliminating experimental mortality as a rival explanation for cross-sectional differences, it is appropriate to ask what, if any, effects this factor could possibly have on the estimation of impact in this study? The evidence suggests that most arguments that come to mind are in the wrong direction. For example, it might be assumed that those who drop out will be less healthy because they are likely to be less conscientious about their health and medical care than those who stay. In fact, as was seen in Table 5, the opposite is true: the experimental group is less healthy at initial screen than the control group.

There does not seem to be a feasible way to control for experimental mortality in the cross-sectional (02–04) comparison. We can, however, again use the inferential feature of this research design to combine the results of cross-sectional and longitudinal comparisons. If the longitudinal (01–02) comparison, which is unaffected by experimental mortality, discloses the same impact after all other threats have been controlled, then there would be no empirical evidence that the EPSDT program impact hypothesis is less plausible. To the extent that they differ, experimental mortality may be the source.

Findings

Age Adjustment. Table 6 shows fairly consistent increases in 01 and 04 ARs when the age adjustment is made. The difference between each group's crude and standardized health status must be attributed to the differences between age distributions compared with the two-screen group (02) at rescreening; for the 01 group, there is about 8 per cent difference between its unstandardized and standardized HS (2.61 per cent versus 2.83 per cent) and 6 per cent difference for 04 (2.58 versus 2.74 per cent). In both comparisons, the effect of age standardization is to improve the experimental (02) group's health status relative to those of the comparison groups. In other words, compared with children at rescreening (02), their own health status at first screen (01) and that of a control group being screened for the first time (04), they appear to be slightly worse due (first) to the favorable age distributions of the first-screen groups.

Longitudinal Adjustment. After the age distributions of the control groups have been standardized, further adjustments are needed: observed changes (01*–02) must be adjusted to control for the effects of increased screening rigor, and cross-sectional differences (02–04*) must be ad-

TABLE 6. Comparison of Abnormality Rates for 02 with 01 and 04 Before and After Age-Adjustment

DA	AR%					
	02	01	→ 01*	04	→ 04*	
HS	2.66	2.61	→ 2.83	2.58	→ 2.74	
Medical	1.98	1.58	→ 1.55	1.96	→ 1.82	
Vision	11.95	12.00	→ 13.13	9.96	→ 11.01	
Hearing	2.15	1.68	→ 1.79	1.99	→ 2.09	
Dental	7.13	7.57	→ 9.79	6.26	→ 7.58	
Laboratroy	1.06	0.94	→ 0.85	1.25	→ 1.08	
Behavior	0.63	0.71	→ 0.70	1.00	→ 1.05	

* An age-adjusted rate.

justed to control for the effects of self-selection for rescreening. Differences that are observed in both the longitudinal and cross-sectional comparisons of adjusted abnormality rates may then be interpreted as evidence of program impact.

The adjustment procedure applied to instrument changes is simply to calculate the per cent change from 01* to 02 and then to subtract from that the per cent change from 03 to 04, due to instrument/history change. The adjustment factors are based on estimates from Table 4 and are signed positive if more abnormalities are found in 1979 than in 1977, and negative if fewer abnormalities are found (Table 7). Similarly, the signs of the unadjusted and adjusted per cent changes indicate the direction of change in ARs before and after adjusting for instrument effect—a negative change indicates reduced abnormalities.

Before the adjustment for instrument effect, there is evidence (Table 6) to support claims of some improvement in health status between age-comparable screen and rescreen results across time. When we adjusted for the increased rigor of the 1979 screening protocol, compared with that in 1977, we find the abnormality rate in 1979 to be about 27 per cent lower. Thus, in terms of the original HS index for the two-screen group in 1977, an average of 26.1 abnormalities were found per 1,000 tests; for the same children in 1979, 19.0 abnormalities per 1,000 tests were found.

Table 7 also allows us to estimate relative impact in different diagnostic areas, although it should be kept in mind that these estimates are based on differing numbers of tests and are therefore of variable reliability. Even after adjusting for age and instrument changes, the greatest improvement appears in the behavioral area, with an almost 60 per cent lower AR.

Cross-Sectional Adjustments. Since cross-sectional comparisons of rescreen results (02) with those of an age-adjusted control group (04*) are subject to selection/regression effects, the per cent differences in rates between 02 and 04* are adjusted in Table 8 based on estimates from the 01–03′ comparisons in Table 5. The sign of the adjustment factor is positive for all DAs, indicating that the two-screen sample had on average higher abnormality rates on first screen than others receiving first screens at the same time. At this point we must recognize that having a significantly sicker group of children for the two-screen sample (at time 1) could mean that the ob-

* Age distributions of control groups have been standardized.

TABLE 7. Longitudinal Changes Adjusted for Instrument/History Effects

DA	Per Cent		
	01*–02 Change	03–04 Adjustment Factor	Adjusted 01*–02 Change
HS	− 6.01	+21.13†	−27.14
Medical	+27.74	+50.77†	−23.03
Vision	− 8.99	+10.06	−19.05
Hearing	+20.11	+44.20	−24.09
Dental	−27.17	−10.95	−16.22
Laboratory	+24.71	+43.68†	−18.97
Behavior	−10.00	+49.25†	−59.25

* An age-adjusted rate.
† Adjustment factors based on significant 03–04 differences ($p < 0.05$). (See Table 4.)

served differences between 02 and 04* are at least partly due to statistical regression in spite of the control. Campbell and Stanley suggest, however, that in the case in which the experimental group is self-selected—such as the two-screen sample—the problem of uniform regression effects between experimental and control group becomes less likely, but interaction effects are more likely.[3]

The adjustment procedure summarized in Table 8 parallels that used for the longitudinal comparison. In this case the per cent AR difference between 01 and 03′ is subtracted from the per cent difference between 04* and 02. As in the longitudinal comparison, the negative signs of several 04*–02 differences, including the overall HS, indicate that the rescreened children are slightly healthier than those of similar age just entering the system in 1979. The estimated adjustment for selection/regression results in the rescreened group showing about 29 per cent fewer abnormalities than the age-adjusted comparison group at first screen.

Conclusions. The similarity of results of the two modes of comparison is more that accidental: the rescreened group has between 27 and 29 per cent lower overall abnormality rate, whether compared with itself over time or with a control group.

It is still possible for either one of these differences in isolation to be explained by rival hypotheses. The longitudinal difference could be due to uncontrolled statistical regression or selection–maturation (and other) interaction effects; the cross-sectional difference may be attributable to experimental mortality. But the plausible rival hypotheses are different in the two comparisons. The research design included this inferential feature of changing contexts deliberately to enable us to isolate and gauge the impact of the EPSDT program based on secondary data analysis.

Summary

The findings presented here attest to the beneficial effect of EPSDT on the health status of children served. Although narrowly focused, our conclusion is definitive and unequivocal: periodic screening is associated with a decrease in the prevalence of abnormalities requiring care.

The results of this study do not permit a definitive statement on the ultimate impact

* Age distributions of control groups have been standardized.

TABLE 8. Cross-Sectional Differences Adjusted for Selection/Regression Effects

DA	Per Cent		
	04*–02 Difference	01–03′ Adjustment Factor	Adjusted 04*–03 Difference
HS	− 2.92	+26.09†	−29.01
Medical	+ 8.79	+24.41†	−15.62
Vision	+ 8.54	+32.56†	−24.02
Hearing	+ 2.87	+29.23	−26.36
Dental	− 5.94	+27.18†	−33.12
	− 1.85	+11.90	−13.75
Behavior	−40.00	+ 4.41	−44.41

* An age-adjusted rate.
† Adjustment factors based on significant 01–03′ differences ($p < 0.05$). (See Table 5.)

of EPSDT and certainly cannot be construed as a blanket endorsement of screening programs in general. We have, on the contrary, deliberately restricted the operational definition of "impact." Our concern is with health status impact, and not with such equally important outcome measures as equity, cost–benefit or client satisfaction. Furthermore, the term "health status" is used exclusively in terms of abnormality rates and does not measure level of functioning, risk status or any of the other significant dimensions of well-being. Health status (and, ultimately, impact) can be measured only by isolating a set of very specific variables on which data are available and from which valid conclusions may be drawn. Consequently, the results of this analysis cannot be interpreted as an estimate of the only, or even of the most important, impact of the EPSDT program. Also, these results do not permit an evaluation of the extent to which EPSDT has achieved its ultimate goal of decreasing overall dependency in the target population. In fact, such an evaluation may never be possible for a program with ultimate objectives as complex and comprehensive as those of EPSDT. Recognition of these limitations in no way diminishes the significance of specific impact that has been identified and documented.

The credibility of any impact estimate depends to a great extent on the power of the research design to eliminate competing explanations for the observed phenomena. Thus, the distinctive methodology employed in this study suggests policy implications that are no less important than those suggested by the substantive conclusions. Specifically, the value and feasibility of secondary data analysis as a method for evaluating program impact has been convincingly demonstrated.

However, it is important to recognize that such secondary analysis is not easily accomplished. Our initial attempts to measure impact produced ambiguous results—no significant change in risk status occurred between first screen and rescreen, despite clear evidence of problem resolution. The actual changes in health status were revealed only after a complex series of mathematical adjustments.

In spite of its complexity, the research design employed here offers several distinct advantages. Using a quasi-experimental design, we can estimate impact without incurring the cost of primary data collection; in addition, such a design avoids the ethical difficulties of selective denial of treatment. Because it is both cost-effective and unobtrusive, the methodology used in this study may be applicable to other operating environments. Through the use of similar secondary data analysis, a national EPSDT monitoring system could be developed, and the impact of screening on the health status of children in a variety of environments could be estimated.

EPSDT, like other public service programs, presently finds itself in an environment of both scarce and shrinking resources and a constant or increasing demand. Estimating the impact of such a program on those it serves is therefore an essential first step toward improving the efficiency of that program by maintaining or increasing the level of services but at a lower cost.

References

1. Shadish WR. Effectiveness of preventive child health care. HCFA Grants and Contracts Reports publication no. 03099, April 1981.

2. Lieberson S, Hansen LK. National development, mother tongue diversity, and the comparative study of nations. American Sociological Review 1974;39:523.

3. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1963.

4. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand-McNally, 1979.

5. Mueller JH, Schuessler KF, Costner HL. Statistical reasoning in sociology. Boston: Houghton-Mifflin Company, 1977.